Acronym	**HydroGen**
Challenge	**Information & Communication Society**
Year	**2014**
Duration	**42 months**

Comparative Metagenomic for Measuring Biodiversity, Application to Ocean Life Studies

Table of contents

Abstract

The HydroGen project aims to design new statistical and computational tools to measure and analyze biodiversity through comparative metagenomic approaches. The support application is the study of ocean biodiversity based on the analysis of seawater samples available from the Tara Oceans expedition.

Comparative metagenomic is a new field aiming at providing high-level information based on DNA material extracted and sequenced from different environments. The problem is not to taxonomically identify the various living organisms present in the various environments. The purpose is mainly to estimate proximity between two or more environmental sites at the genomic level.

One way to estimate similarity is to count the number of similar DNA fragments. The sequencing of a single environment generates a dataset of 10^8 to 10^9 short DNA sequences ranging typically from 100 to 150 base pairs (called reads). From a computational point of view, the problem is thus to calculate the intersections between datasets of reads. To evaluate this similarity, the traditional way is to compute a score attached to an alignment between two reads.

The main drawback of this technique is that the number of alignments to compute is excessive (10^{16} to 10^{18} between 2 samples). Furthermore, if several hundreds of metagenomic samples are involved, then this approach is currently not achievable with current alignment techniques. The main challenge of the HydroGen project is to propose alternative methodologies to efficiently compare such volume of metagenomic samples.

The validation of our methodologies, and the scaling of algorithmic and statistical tools developed during the project, will be done from environmental questions linked to the study of the biodiversity of oceans. The Tara Oceans expedition has collected hundreds of seawater samples that are currently sequenced. Hundred of metagenomic data sets are thus available to

the scientific community. This mass of data will be used as the primary material in the framework of the HydroGen project.
The HydroGen project gathers 3 research teams with complementary competences in algorithmic, statistics and genomics: INRIA-GenScale, INRA (MIA/AgroParisTech+MIG) and CEA-CNS-LABIS.

People

Partner	Family Name	First Name	Current Position	ETP (nb months)	Role and Responsibility in the HydroGen project
INRIA	Lavenier	Dominique	DR CNRS	14	Coordinator NGS Algorithm design
INRIA	Peterlongo	Pierre	CR INRIA	11	Leader task 2 Metagenomic algorithm
INRIA	Lemaitre	Claire	CR INRIA	9	Algorithm/soft development
INRIA	Deltel	Charles	IR INRIA	8	Software development
INRIA	PhD		*To be recruited*	36	Algorithm (task 2) & validation (task 4)
INRA	Robin	Stéphane	DR INRA	6	Leader task 3 Statistical modeling
INRA	Schbath	Sophie	DR INRA	6	Statistics of k-mers
INRA	Mariadassou	Mahendra	CR INRA	6	Statistical modeling, sparse PCA
INRA	Chiquet	Julien	MC Evry	6	Statistical modeling, sparse PCA
INRA	Aubert	Julie	IE INRA	4	Statistical modeling
INRA	Ouadah	Sarah	MC APT	6	Statistical modeling
INRA	Lebarbier	Emilie	MC APT	4	Statistical modeling, mixture models
INRA	Post-doc		*To be recruited*	24	Statistical methodology (task 3)
CEA	Jaillon	Olivier	CEA Researcher	8	Leader task 4 Metagenomic sequence analysis
CEA	Aury	Jean Marc	CEA Engineer	8	Sequence Analysis
CEA	Pelletier	Eric	CEA Researcher	6	Genomic Analysis
CEA	Wincker	Patrick	CEA Researcher	3	Genomic Analysis
CEA	Wessner	Marc	CEA Engineer	3	NGS data analysis
CEA	Enginner		*To be recruited*	12	Evaluation of sequencing technologies for metagenomic

1. Context, positioning and objectives

Context

Comparative metagenomic is a new field aiming to provide high-level information based on DNA material extracted and sequenced from different environments. More precisely, the problem is to determine how similar environments and living organisms are from direct DNA or RNA analysis. At this level, the problem is not to taxonomically identify the various living organisms present in the various environments. The purpose is mainly to estimate proximity between two or more environmental sites at the genomic level.

One way to estimate similarity is to count the number of similar DNA fragments. The sequencing of a single environment generates a dataset of 10^8 to 10^9 short DNA sequences ranging typically from 100 to 150 base pairs (called reads). From a computational point of view, the problem is thus to calculate intersection between datasets of reads. Two reads from two different datasets will be considered to belong to the same organism (or species) if they exhibit a *strong* similarity. The usual way to evaluate this similarity is to compute an alignment. The score attached to this alignment indicates the degree of similarity.

This method has already been experienced with data of Tara Oceans (at a very low scale) and has demonstrated the ability of characterizing diversity by analyzing a series of DNA samples from Mediterranean plankton. It has been observed that raw read comparison is well suited in the context of a multi-site survey to rapidly provide an initial assessment of an ecosystem. In particular, other methods require to assemble reads for reducing the volume and the data complexity, but this leads to restrict the survey to only 2% of the sequences in such complex samples. On the contrary, raw read comparison strategy gives to genomic experts an overview with *no a priori* of the total sampled genomic diversity.

However, the main drawback of this technique lies in the way intersections are computed. It requires processing an extremely huge number of alignments. Two datasets of N reads implies the computation of N^2 alignments. For $N = 10^8$, for example, the computing power required by BLAST (the main software of the domain) to perform such computation will need a full week with a 5000-node computer.

Processing a few metagenomic samples that way is thus achievable, but does not provide realistic approach for large-scale metagenomic projects. The Tara Oceans expedition, for instance, has already collected hundreds of seawater samples coming from all the world's oceans. Establishing a global world heat map requires computing millions of intersections. This is currently unattainable using BLAST-like approaches.

Objectives

A first objective of the HydroGen project is to tackle this challenge by investigating new computational methodologies able to process this huge volume of information. Metagenomic datasets represent incomplete and partial information, and accurate metrics such as similarity score provided by alignment can probably be revisited to be adapted to comparative metagenomic. Simpler metrics would reduce computational complexity, and then execution time.

A complementary approach is also to perform statistical analysis on the metagenomic datasets before running time-consuming treatments. Similarity estimation can be conducted by analyzing, for example, a representative set of k-mers (short sub-sequences issued from reads) sampled from the datasets. A relevant statistical analysis of the complete datasets may give an

informative overview of the global similarity and provide clues to focus on specific datasets regarding a specific biological problem.

Statistical analysis can also be used to efficiently enhance computational steps, especially on the elementary task of processing two metagenomic datasets. Depending on the intra diversity of a dataset, processing all reads is probably not necessary. A judicious sampling could significantly limit the amount of computation while providing equivalent result compared to a systematic read comparison. The problem is how to estimate the intra diversity of a metagenomic dataset.

The second objective of the HydroGen project is thus to develop statistical tools adapted to comparative metagenomic. First, by providing a quick overview of the biodiversity between a large set (several thousands) of metagenomic samples. Second, by providing relevant statistical indicators allowing computational steps to be fasten.

Developing tools and methodologies for comparative metagenomic cannot be disconnected from real biological or environmental questions. In the HydroGen project we propose to associate an important application component linked to the study of ocean biodiversity through metagenomic analysis. This last part will act as a test bench (1) to validate our methodology and (2) to demonstrate that global metagenomic approaches can bring supplementary information to answer biological or environmental questions.

To summarize, the goal of the HydroGen project is to develop efficient methodologies combining statistical and algorithmic approaches and targeting ambitious comparative metagenomic studies.
HydroGen will particularly focus on the study of ocean biodiversity based on metagenomic data from the Tara Oceans expedition.

It is important to note that methodologies and tools developed during the HydroGen project do not restrict their exploitation to ocean studies. Any kind of metagenomic projects involving the processing of a large number of metagenomic samples will benefit of the outcomes of HydroGen. In the present case, the availability of the Tara Ocean data represents a great opportunity that perfectly fits with HydroGen objectives.

State of the art

Comparative metagenomic

Comparative metagenomics aims to compare together several metagenomes. Comparing two or more metagenomes is an efficient way to understand how genomic differences of communities affect the physico-chemical factors of an ecosystem. For example, to study soil composition, Ishii et. al. have compared various metagenomes implied in the denitrification (Ishii, 2009). Complementary, this approach also tells us the way environment infers on metagenomes (Wooley, 2010). For instance, it has been shown that nickel-rich environment promotes the development of resistance genes compared to a specific control environment (Mirete, 2007).

Furthermore, comparative metagenomic can be used to explore different metagenomic samples simply based on their contents, i.e. without trying to identify the contents. The “global ocean sampling“ (GOS) expedition is a good illustration of this approach: it efficiently tested this strategy by clustering tenth of samples by a comparison analysis of their contents. Results showed that metagenomic samples from geographically close locations or sharing common environmental factors tend to cluster together (Venter-2, 2007).

Comparative metagenomic can be supported by several strategies. One is to use quantitative or functional metagenomic information of specific sequences. 16S-RNA, for example, are good candidates to differentiate metagenomic contents (Jaenicke, 2011). The MG-RAST software aims to perform comparative taxonomic analysis (Port, 2012), MEGAN uses the abundance information for clustering taxons (Shakya, 2013), and IMG/M studies parental links or different ecosystems (Cardoso, 2012). All these methodologies exploit current knowledge but leave aside a large part of the information contained in all the other sequences of a metagenome.

To extract more information from *a priori* unknown sequences, various methods have been investigated. One can explore the GC content of the metagenomes. As a matter of fact, between different genomes, the GC content is not uniform, and the study of Foerstner demonstrated that the environment has a strong impact on the GC content of the organisms (Foerstner, 2005). The analysis of the GC content can thus be a way to measure a similarity between metagenomic samples, or at least be a differentiator element.

Another method analyzes a set of genetic markers. Raes et al. took 26 of them considered as essential for the cell survival (Raes, 2007). These selected markers evolve very slowly and are linked to important functions inside the cell. Furthermore, they are preserved throughout the living organisms and appear only once per genome. From the quantity of the markers present in the metagenomes, the average size of the genomes is estimated. Even in complex metagenomes, the total number of markers is directly proportional to the number of genomes. The marker density (# markers / # genomes) is thus inversely correlated to the average size of the genomes present in the metagenome. This ratio can serve as a measure to compare different metagenomes.

Metagenomic projects

Before presenting various techniques for extracting knowledge from metagenomic datasets, we first give a short overview of typical metagenomic projects. The first one (MetaHIT) deals with human health, the second one (METASOIL) with environmental soil study, and the third one (GOS) with ocean biodiversity. These three projects target different biological problems. They have in common to use NGS metagenomic datasets as input from which new knowledge are extracted. The purpose is not to describe in detail the bioinformatics methodology developed by these projects, but to give a flavor of what can be expected from these new approaches, and to show the diversity of biological questions that can be addressed.

MetaHIT (Metagenomic of Human Intestinal Tract) aims to explore the human microbiome and to decipher relationship between microorganisms and human health (Ehrlich et al. 2010). MetaHIT focuses on two specific diseases: obesity and inflammatory bowel. The three objectives of the project were: (1) to provide a complete catalog of microbial genes present in the Human intestine; (2) to determine from this catalog which genes are present in a cohort of sick and healthy people, and how often; (3) to establish relationship between genes and diseases.

Metagenomic samples of 124 individuals from different origins have been analyzed (540 Gb of DNA sequences), leading to a catalog of 3.3 millions of bacterial genes. The comparison of each metagenome to this catalog has shown the existence of 3 enterotypes that could be linked to people diet. Researchers also inventory about 1000 different bacterial species present in Human gut. However, a specific individual have only an average of 160 species.

Over 19000 different functions have been identified in the gene catalog. The statistical analysis indicates that all of the functions are present in the 124 samples giving an exhaustive view of the genetic potential of the bacteria from the human gut. A large proportion of the functions, over 5000, were never found before. This illustrates the novelty of the metagenomic approach.

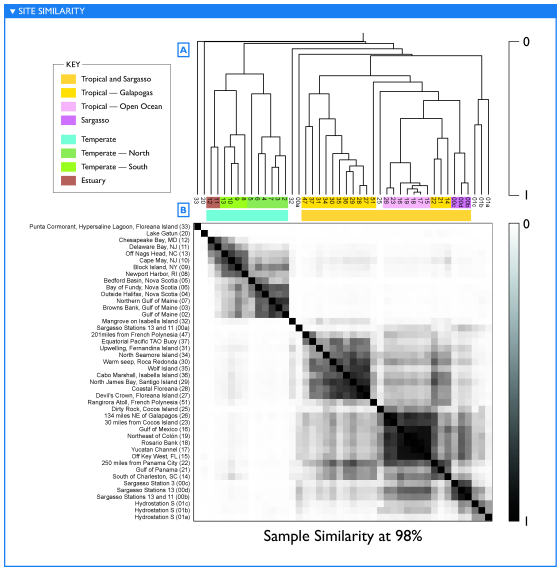
METASOIL (Metagenomic discovery and exploitation of the soil microbial community) is an ANR project (2009-2013) with several objectives: to create a metagenomic and genetic resource of an internationally unique, well characterized soil; to identify and contrast the diversity of the active bacteria and their key functional genes, establishing a baseline of a "healthy" soil; to correlate variation in genetic diversity of soil microbial ecosystems to soil management and environmental perturbations (climate change, pollution) over the last 100 years.

Samples were first processed with the MG-RAST tools to extract gene functions. 835 were identified and used for downstream studies. One of them considers 13 metagenomes from the Park Grass Environment of Rothamsted, England, two metagenomes from other places and one seawater metagenome. These metagenomic samples were studied together to evaluate various criteria, such as sequencing technology biases, soil depth of the sampling or seasons of the sampling. An interesting point highlighted by the researchers is that much more knowledge can be extracted from metagenomic samples coming from various places compared to an analysis focusing on functions expressed in only one metagenome (Delmont 2012)

GOS (Global Ocean Sampling) project is an expedition funded by J.C. Venter laboratory to study ocean biodiversity (Venter-2 2007). Seawater samples from 41 different places have been collected, leading to the effective sequencing of 44 metagenomes. These samples have been used to answer several biological questions and to perform various analyses. One of them is the creation of a *heatmap* to give an overview of the ocean diversity (Venter-1 2007). This map has been built based on the DNA content of the datasets, i.e. directly by comparing the DNA sequences. The resulting heatmap clearly shows a strong correlation between geographically close samples.

Similarity between samples in terms of shared genomic content (Venter-2, 2007)

Genomic similarity is an estimate of the amount of the genetic material in two samples that is the same at a given percent identity cutoff—not the amount of sequence in common in a finite dataset, but rather in the total set of organisms present on each filter (A) Hierarchical clustering of samples based on pairwise similarities. (B) Pairwise similarities between samples, represented as a symmetric matrix of gray scale intensities; a darker cell in the matrix indicates greater similarity between the samples corresponding to the row and column, with row and column ordering as in (A). Groupings of similar filters appear as sub-trees in (A) and as squares consisting of two or more adjacent rows and columns with darker shading.



J.C. Venter-1 et al., The sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through eastern tropical pacific. Plos Biol, 5(3):e77, Jan. 2007

The overall comparison process was performed by first computing a database of overlapping DNA sequences coming from the 44 metagenomic samples. 1.2 billions of overlaps were identified and then used to construct the heatmap. The raw comparison was performed with a modified component of Celera tool to detect first overlapping of 14 identical nucleotides. Then, a seed extension strategy is applied, and only alignments above a given identity percentage are retained. This standard comparison strategy was possible here since the volume of information

to process was not so high: a GOS seawater sample contains an average of 175 K sequences, of ~ 1250 bp (220 Mbp, Sanger technology). By comparison, the size of a Tara Oceans sample is about 150 Gbp.

Sequence similarity tools

One of the most effective and direct way to globally compare metagenomes lies in the availability to aggregate local similarities from the primary sequences generated by NGS machines. The problem is thus to compare each sequence from one dataset to each sequence of another one, and to determine a global score (from local ones) that expresses an overall distance (or similarity) between metagenomes. The basic tools available for such a task are those dedicated to the computation of alignments such as BLAST, KLAST, BLAT or USEARCH (non exhaustive list). Each of these software includes powerful heuristics to speed-up the computation, compared to the optimal dynamic programming algorithm.

BLAST is based on the seed-extend heuristic: an anchor (seed) is first searched between two sequences, then left and right extended if possible (Altschul, 1997). The size of the seed determines both the sensitivity and the speed of the computation: longer the seed, shorter the execution time (but smaller the sensitivity). BLAST has been primarily designed to query genomic databank and is thus not optimized to perform intensive comparison between banks.

KLAST follows the BLAST strategy (seed-extend heuristics) but has been explicitly designed for genomic bank comparison (Nguyen, 2009). The 2 banks are fully seed-based indexed into the memory allowing computation to be dispatched independently on multicore processors. Sensitivity can also be tuned through seed manipulation: to reduce the execution time, only a fraction of the seeds can be considered. Compared to BLAST an average speed-up of 10 is observed (same sensitivity).

BLAT on DNA sequences has been designed to find sequence of 95% and greater similarity (Kent, 2002). It is used to compare a set of sequences to a complete genome, but can be diverted for metagenomic purpose. It keeps an index of partially overlapped seeds of the entire genome in memory. This partial seed overlap has a huge impact on the execution time. Speed-up of two orders of magnitude higher than BLAST is achieved. On the other hand, only high similarities are reported.

USEARCH explores another heuristics: it counts the number of identical short words in common between two sequences and decides whether it is worth or not to continue the alignment computation according to a threshold value (Edgar 2010). The execution time is somewhere between BLAST and KLAST with comparable sensitivity.

Specific tools for comparative metagenomic

These four tools are not specific to metagenomic and don't scale well for very large datasets where the number of sequences to process is high (from 10^8 to 10^9 100-150bp reads). Limitations come from the excessive computation time (even for tuning sensitivity to a minimal threshold) or from the size of the index structure for software that index databank in memory. Recently, a few tools have been proposed to overcome these limitations: crAss, TriageTools and Compareads.

crAss compares several metagenome by first considering a "soup" of all metagenomes. This soup is "assembled" to build contigs with a classical assembly tool (Dutilh, 2012). These contigs can potentially be shared by genomes of different metagenomes. The following step gives contigs and raw sequencing data as input to crAss. After mapping the reads on the contigs, crAss counts the number of sequences of each metagenome that match contigs. This is a way to

determine if contigs are shared between several metagenomes and to measure the relationship between metagenomes.

TriageTools actually has a larger scope than metagenomic (Firmeli, 2013). This is a sort of NGS toolkit for partitioning and prioritizing analysis of high-throughput sequencing data. Based on k-mer similarity considerations, the tool performs extraction of reads from a dataset based on targeted sequences located in another dataset. The targeted sequences are split into k-mers and indexed in a hash-table. Comparison with a read dataset is done by querying the hash-table: for a given read, if its k-mer composition overcomes a threshold value, then the read is selected. We can see here, that the notion of similarity is much less stringent than similarity issued from score alignments. Similarity is simply provided by a k-mer count information.

Compareads works on the same model (Maillet, 2012). But, contrary to TriageTools, which differentiates targeted sequences of other data, this software processes two datasets with no distinction. Its only goal is to provide a global similarity by selecting all common reads between two metagenomic datasets. The basic idea to find such reads is to assume that they share common k-mers. An index, based on Bloom filter, is built to compress the k-mer composition of one dataset into memory. The selecting process is similar to the previous tool.

The k-mer count method developed by TriaTools and Compareads are currently the only ones able to tackle large genomic datasets in a reasonable amount of time, and with a reasonable memory footprint: for example, two datasets of 10^8 reads can be processed in less than 10 hours with a memory not exceeding 8GBytes.

Compareads is currently developed by the INRIA/GenScale research team with strong algorithmic and optimized data structure orientations. Even if several gaps of magnitude have been achieved, from the execution time point of view, further significant improvements won't be possible without adding mathematical and statistical elements aiming at limiting the computational complexity. The next section gives a brief overview of the domain.

Statistical sequence analysis

Statistical analysis of sequence k-mer frequencies is now well established in the literature, in particular the question related to find k-mers with exceptional frequencies, i.e. significantly over- or under-represented in a given sequence. The k-mer count distribution has been extensively studied in random sequences (under Markovian models) leading to exact but more commonly used approximated distributions depending on the sequence length and k-mer characteristics (Gaussian, binomial, Poisson, compound Poisson etc.); see for instance the book written by two members of the HydroGen project (Robin, Rodolphe, Schbath, 2005). We can then easily imagine to use the R'MES software developed by the INRA research team (Schbath and Hoebeke, 2011) to identify over-represented k-mers in each metagenomic dataset. Such exceptional k-mers could then be used to directly compare metagenomic samples but also to limit Compareads to the search of reads that share common exceptional k-mers.

We have seen previously the limitation of sequence similarity tools for the metagenomic purpose tackled in this project. Several alignment-free sequence comparison methods have been proposed in the past (see for instance Song et al. 2013 and references therein). Most of them are based on the sequence k-mer compositions, and they usually differ from the way distances between these k-mer count vectors are computed (Dai et al, 2008). Very recently, Wang et al. applied some of them (e.g. the D2 statistics) to compare meta-transcriptomic samples (Wang et al, 2014). However, all these methods neglect the fact that k-mer occurrences overlap in the sequence leading to intrinsic dependency between the counts. Nevertheless, the covariance matrix between the counts can be explicitly characterized (Reinert, Schbath, Waterman 2005) and should be taken into account.

2. Scientific and technical program, project organization

Overall organization

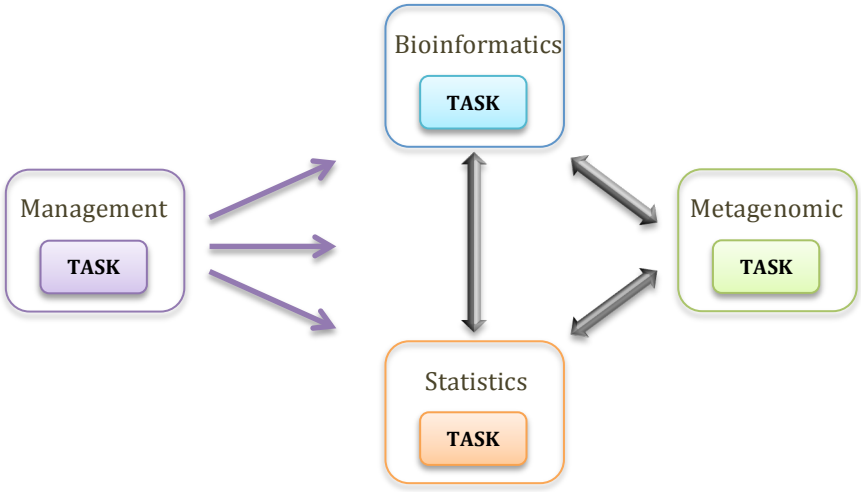
The study will be conducted in the frame of the Tara ocean project. The great advantage is that a large number of seawater metagenomic samples have been collected, many of them have already been sequenced, and that the complete set should be sequenced in the two next years, leading to Tera Bytes of data to analyze.

The HydroGen project will be structured around three main axes:

1. The development of comparative metagenomic bioinformatics tools strongly linked to axe 2, to power the analysis of metagenomic datasets.
2. The development of statistical methodologies to provide 1) quick glance of biodiversity among the whole metagenomic dataset; 2) statistical indicators for driving the computational step.
3. The study of environmental applications through metagenomic sequences.

These 3 axes lead to naturally organize the HydroGen project into 4 main tasks that are highly interconnected together:

- Task1: Management
- Task2: Bioinformatics tools
- Task3: Statistical methodology
- Task4: Analysis of metagenomic sequences



Task 1 is devoted to the management of the HydroGen project. Task2 aims to develop new bioinformatics tools dedicated to comparative metagenomic. Task 3 focuses on statistical tools that will guide algorithms to generate relevant information. Task 4 will apply both bioinformatics and statistical methods developed in task 2 and 3 to analysis Tara Ocean data. Feedback from task 4 to task 2 and 3, regarding the quality of the analysis, the relevance of the produced information, the correctness of the results compared to established references will help to steadily improve the strategies, the methodologies and the tools. Task 4 will also study the impact of different metagenomic sequencing strategies, including new emerging technologies such as Oxford Nanoport Technology or Illumina Moleculo Technology.

Task 1: Management

Task leader: D. Lavenier

Objectives: to induce and nurture synergies between partners towards the completion of the HydroGen project objectives, with the further goal of setting long-term collaborations; to make sure that the HydroGen commitments to ANR are fulfilled in due time; to promote the research activity of the project.

Working program:

Organizing the synergies inside the HydroGen project

Synergies and interactions between partners will be promoted and strengthened through physical meeting or video-conferencing to join forces, at the individual level, to crack a particular problem proposed by either a task coordinator or individuals themselves, and finally to keep track of the progress by a tight monitoring. This is essential to the success of the project as we come from different communities and cultures.

In addition "official" general meetings will be organized with all participants at the dates below to globally synchronize (or possibly re-synchronize) task interactions, or re-schedule, if needed, subtasks according to the progress of the project. The final meeting will be open to a large audience to present results of the HydroGen project.

- T0 Kickoff meeting - Rennes
- T0+15 Synchronization meeting - Paris
- T0+30 Synchronization meeting - Rennes
- T0+42 Final meeting - Paris

Promoting the HydroGen project

A web site will be created with the following objectives:

- Presentation of the HydroGen project
- Availability of intermediate results, publications, etc.
- Space work for the partners with intranet access
- Depository for software developed during the project

Deliverables:

- D1.1 ANR 6 month intermediate reports
- D1.2 HydroGen web site
- D1.3 ANR mid-project report
- D1.4 ANR final report

Risks: if for some reasons the coordinator fails in doing its management work, a replacement manager will be found among the tasks leaders.

Requested resources:

	Mission	Justification
INRIA	3.0 K€	2 missions to Paris for 4 people (T0+15, T0+42 meetings)
INRA	6.0 K€	2 missions to Rennes for 8 people (T0, T0+30 meetings)
CEA	3.0 K€	2 missions to Rennes for 4 people (T0, T0+30 meetings)

Task 2: Bioinformatics tools

Task leader: P. Peterlongo

Objectives:

The challenge is to develop bioinformatics tools taking as input raw metagenomic datasets, and providing as output an estimation of the similarity between two environments. Exploratory researches inside the INRIA/IRISA GenScale group investigates the following strategy:

- **K-mer based metric**: to estimate proximity between two reads, an approximate similarity based on common k-mers between reads is considered. The great advantage of extracting similarity that way is that the computational complexity is drastically reduced.

- **Probabilistic data structures**: Reads from metagenomic datasets are split into k-mers that are indexed into Bloom filter structures. Computing intersections between datasets is performed by querying these data structures.

Ongoing experimentations inside GenScale demonstrate that the intersection between two datasets of 10^8 reads can be processed in a few hours. They also show that heat maps computed with that strategy are similar with heat maps generated from blast-like methods. However, this promising strategy still needs further investigation to address very large metagenomic projects involving thousands of metagenomic datasets. Today, the method focuses on the comparison of two metagenomic samples. It needs to be extended to the comparison of N-to-N metagenomic datasets. Based on this exploratory work, a first objective of task 2 is thus to scale this methodology to ambitious metagenomic projects, such as Tara Oceans, for processing hundreds or thousands of metagenomic samples generated by the high throughput sequencing projects (subtask 2.1).

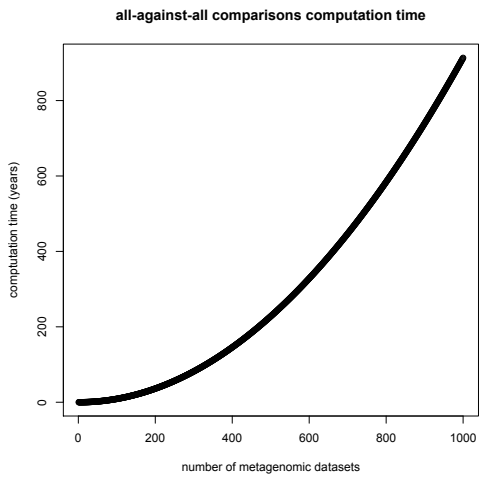
Another interesting challenge is to query such huge set of metagenomic samples in the same way biologists query genomic databases through BLAST server: Given a DNA sequence, or a small set of DNA sequences related to a specific function, return a list of metagenomic sample where these sequences significantly appear. Again, the standard blast-like approach, based on alignment detection, is not well suited for providing result in a reasonable time. The solution is both in the development of efficient data structures and adapted metric as mentioned above, all of this driven by solid statistical methodologies. This problem will be addressed in subtask 2.2.

Working program:

Task 2.1: Comparing large sets of metagenomic samples

As previously mentioned, the GenScale team showed the feasibility of comparing two metagenomic datasets composed of hundreds of millions of reads in a few hours. The aim of this task is to go much further to tackle the processing of a large number of datasets.

Current approaches are restricted to 1-to-1 dataset comparisons. Computing N-to-N datasets requires a quadratic number of comparisons. As shown on the figure below, the current state-of-the-art methods do not scale. For example, processing all Tara Oceans metagenomic datasets in that way would require centuries of computation.



Scaling to N-to-N metagenomic dataset comparisons is an ambitious task that requires removing critical bottlenecks such as computation time, space memory, disk storage, and the non-uniformity of datasets (qualitatively and quantitatively). The task success mainly lies on the synergist combination of algorithmic, statistical and biological skills for exerting a leverage effect to focus on informative data and to process them quickly and efficiently.

The cornerstone is probably to reduce the dataset complexity by applying advanced statistical methods while keeping biologically relevant information. Datasets will be studied at the k-mer level, meaning that full read information is lost. However, this representation presents non-negligible advantages for statistical analysis, redundancy compression, RAM memory storage and indexing techniques.

The envisioned approach consists in considering each sequencing metagenomic dataset as vector of k-mer frequencies. The idea is to propose a reduction and a clustering of all datasets based on this representation. The success of this approach is strongly linked to statistical methods that will be developed into task 3, and to innovative algorithms that need to be imagined. Indeed, the simple counting and representation of billions of k-mers through hundreds or thousands of distinct datasets is challenging and will require the adaptation of counting software such as DSK (G. Rizk 2013) that were initially designed for the counting of unique datasets. Moreover, the data reduction and clustering techniques are time consuming. Both fine and coarse-grained parallelization techniques can also help in significantly reducing computation time.

A raw similarity distance between metagenomic datasets (such as the intersection of common reads) can be difficult to interpret, and possibly non-immediately correlated to real biology considerations. Relative distances between datasets would probably give more relevant information. A consequence is that a systematic and controlled error rate could be acceptable, as it does not affect the relative comparisons.

The possibility to authorize a certain level of errors provides some room to investigate the use of advanced data structures which may contain only partial but necessary information. Thus, again in close collaboration with task 3, we plan to tailor k-mer indexing scheme for maintaining an acceptable error rate. Probabilistic data structures such as Bloom filter are good candidates to capture essential information while maintaining a correct error rate. Second, we propose to sub-sample datasets by artificially reducing their sizes, either randomly or focusing on reads containing specific pieces of information such as discriminative k-mers. These two ways of investigation should lead to greatly reduce both the memory footprint and the computation time.

Task 2.2: Querying large metagenomic databank

The Tara Ocean metagenomic databank is constituted of hundreds of sequencing datasets coming from the seawater metagenomic samples. Each dataset houses hundreds of millions of short DNA sequences (called read) of 100 to 150 bp. The problem is to query this databank in a reasonable amount of time for answering questions such as: in which datasets a particular sequence (gene, virus, etc.) appears ?

Fast conventional string comparison approaches, based on seed-extend strategy to localize potential alignments, will fail to generate results in relatively short time. Blast-like approaches imply to systematically scan all the datasets, leading to obvious I/O bottleneck. Getting all the data in memory, together with an index structure, is definitively prohibitive due to memory space required (10 to 100 To for Tara Ocean).

The approach we plan to investigate is to transform each dataset into a condensed representation of the information based on Bloom filter data structure or, possibly, on other advanced probabilistic data structures. This technique has been advantageously used for genomic assembling purpose and allows large datasets of reads to fit into computer memory. A dataset is split into k-mers that form the primary information that is stored inside the Bloom filter. This data structure can thus be queried from the base of k-mers. The response of the Bloom filter generates false positive. The false positive rate can be tuned according to the wanted precision, the nature of the data, the redundancy information inside the dataset, etc. There is a strong link here with Task 3 that should provide statistical tools or methodologies for optimizing both the probabilistic data structures together with the associated query algorithms.

Even if probabilistic data structures provide a very efficient way to drastically compress and query the primary datasets, the high number of samples to consider may still be an obstacle to directly use this approach. It supposes to have all this "abstract" information into memory. A way to decrease the volume of this information is to consider only "relevant" information by selecting the right primary bits of information, that is, a set of pertinent k-mers. Again, strong interaction with task 3 is mandatory to define solid statistical bases such kernels of k-mers.

Finally, for still limiting the amount of data to process, hierarchical approaches will be investigated, allowing to discard possible non relevant datasets or part of datasets. This will be also conducted with disk technology considerations, especially by experimenting the use of SSDs. We can think of a distributed data structure between part in RAM, part in SSD and part in classical hard drive.

The Tara Ocean dataset will actively serve as test-bench. Experimentation of the querying process will be done in tight interaction with task 4. We already dispose of tens of metagenomic datasets. These data will be used in the early statistical and algorithmic design steps for validating methodologies on prototype software. As Tara sequencing will continue during all the life of the HydroGen project, – leading to more and mode data to analyze – the HydroGen project will naturally scale toward a complete analysis of the Tara Oceans data. More precisely, in the context of the biodiversity ocean study, the work of this subtask should help to visualize, for example, if given plankton species (identified from specific DNA sequences) follow a particular stream.

<u>Deliverables</u>:

Metagenomic comparison software (task 2.1)

- D2.1.1 prototype – internal use (T0+24)
- D2.1.2 reliable software - available to the scientific community (T0+42)

Metagenomic querying software (task 2.2)

- D2.2.1 prototype – internal use (T0+36)
- D2.2.2 reliable software – available to the scientific community (T0+42)

PhD Thesis (see below)

- D2.3 Document and defense of the PhD (T0+42)

Risks:

Methods developed by GenScale, and mostly applied to the global metagenomic comparison (subtask 2.1), have proved their efficiency when analyzing two metagenomic samples. These methods have to be enhanced by statistical material and further algorithmic developments to scale up to multi metagenomic sample analysis. This is one of the main challenges of this project. This task is highly interconnected to tasks 3 (statistical methods) and 4 (application and validation) and therefore its success relies on substantial collaborations with other groups of the project. To ensure efficient collaborations we will plan several working meeting in small groups in addition to annual general meetings. For instance, the post-doctoral student hired on task 3 will visit regularly the Genscale team and the PhD student will visit the Genoscope and spend a 3-month internship in the Computational Biology Group at MIT, Boston, USA

Requested resources:

	Operating cost	Justification
INRIA	2.0 K€	Missions to Paris: working meeting, interaction with tasks 3 & 4
	3.0 K€	Internship PhD to MIT, CompBio group, USA
	9.0 K€	International bioinformatics congress,
	4.0 K€	Publications costs
	30.0 K€	subcontracting for software development
	4.0 K€	Computer resources (PhD)
INRA	3.0 K€	Mission to Rennes
CEA	2.0 K€	Mission to Rennes
	Staff cost	Justification
INRIA	120 K€	1 PhD – 3 years

We request the funding of a PhD student to perform computational algorithmic development and to integrate the statistical expertise brought by task 3. Subtasks 2.1 and 2.2 share common algorithms and data structures from which comparing and querying metagenomic tools can be derived. In addition, as mentioned in the previous section, subtask 2.2 can be considered as a refinement of tools developed in task 2.1. The research work will thus cover both subtasks under the guidance of P. Peterlongo and D. Lavenier. The validation step will be done through applications proposed by task 4. A 3-month internship to the Computational Biology Group at MIT, Boston, USA, during the 2nd year is planned to work on the detection and characterization of families of unknown genes (see task 4.3). This group has already strong interactions with partner 3 (Genoscope), especially on this domain.

We also request an extra expertise from a software company to ensure industrial compliance of the software that will be developed during the project. In addition to consulting, the involvement of this company in the HydroGen project will provide a practical help in the optimization of the software by moving research software prototypes into reliable software, and by taking in charge all mandatory software engineering aspects that is required when software are made available for the international scientific community.

Task 3: Statistical methodology

Task leader: Stéphane Robin

Objectives:

In the proposed approach, each metagenomic sample will be characterized by series of k-mer frequencies. Based on this representation, most of the questions addressed in this project can be casted in the general framework of multivariate statistical analysis for which many techniques exist for a long time, including principal component analysis and clustering, to name of few, the probabilistic counterparts of which refer to a Gaussian setting. However, these standard techniques do not account for the specificities of metagenomic data. The two main specificities are (a) that we will be dealing with count data (with a large proportion of zero counts) and (b) that the dimensionality of the data will be huge. Furthermore, due to possible overlaps between k-mers, their frequencies display a strong dependency structure that can be characterized, but that has no biological meaning. This nuisance dependency has to be accounted for in all the analyses to be carried out.

Working program:

Task 3.1: Reduction of the dimension

As k-mers will be considered, the number of descriptors of each metagenomic sample will be very huge. Thus, scarcity is desirable both to make the results more interpretable and to alleviate the computational burden (c.f. task 2). A natural method to select relevant k-mers would be to concatenate reads of a given sample and to find for k-mers significantly over-represented in each metagenomic sample (R'MES tool). For computationally purposes this preliminary step can be performed for moderate values of k (say k=15) and then one could consider longer k-mers containing these exceptional words.

More generally, principal component analysis is one of the most popular approaches to reduce a large data set to a smaller dimension. Convex relaxation techniques have already been proposed in multivariate analysis resulting in sparse principal component analysis, where each observation is associated only to a limited number of components (Zhou & al., 2006). Again, these approaches have been developed in a Gaussian setting and need to be adapted to count data. We will consider this extension in the (multivariate) generalized linear framework, which allows transposing a series of Gaussian models to non-Gaussian distributions. Within this task, we will first define a Poisson version of PCA and then consider its generalization to Negative Binomial as well as its sparse version.

Task 3.2: Sample clustering

A second natural way to perform dimension reduction is to cluster the samples into a small set of clusters. Distance-based have been proposed to perform alignment-free sequence clustering. Most of them are borrowed from standard hierarchical clustering techniques, which need to be adapted to k-mer counts. Indeed, due to possible overlaps between them, k-mer frequencies display a strong dependency structure that can be very accurately characterized, but it has no biological meaning. The definition of new sequence distances accounting for these correlations (inspired from the Mahalanobis distance) will be considered to avoid spurious clusters.

Apart from distance-based, model-based approaches also exist to classify samples into meaningful groups. Although distance-based methods are sometimes more computationally efficient, the model-based approaches (also known as mixture models) allow more flexibility to account for the specificities of sequence data. Taking advantage of the fact that k-mers counts

are closely related to Markov chain models, we will consider mixtures of Markov chains. Such models are known to be sensitive to a large variability of the sequence lengths, but could be well adapted to the clustering of reads, which have all about the same size.

Task 3.3: Metagenomic sample comparison

Due to the huge number of descriptors and in absence of any prior knowledge, a blind sampling of the k-mers can be considered to fasten the comparisons between metagenomic samples or between one query sample and the database. Indeed, the resulting similarity measure will suffer some uncertainty that will need to be evaluated, accounting for the redundancy that exists in each sample. The evaluation of this uncertainty will be useful to define a compromise between computational efficiency and statistical precision. The evaluation of this uncertainty also depends on the correlation between the sampled k-mers frequencies, the structure of which is known. In a second time we will study how this knowledge can be used to make the sampling more efficient.

As for comparison, based on well-established works on motif counts distribution, tests can be defined to assess the significance of the difference observed between several samples. We will use such tests to select the most discriminant k-mers between the samples under study. Then we will consider the multivariate comparison approach, where several k-mers are considered at once. The reference model for this task is multivariate analysis of variance (manova). Again, the knowledge of the covariance structure between the k-mer frequencies will be useful. Following the same way as in task 3.1, a sparse version of the manova model (known as sparse discriminant analysis) will be considered and adapted to the non-Gaussian nature of read counts.

Deliverables:

D3.1 Statistical tools & methodologies for reducing high k-mer dimension (T0+30)
D3.2 Statistical tools & methodologies for metagenomic sample clustering (T0+24)
D3.3 Statistical tools & methodologies for metagenomic sample comparison (T0+36)

Note that the outputs of subtasks 3.1, 3.2 and 3.3 will be used by subtasks 2.1 and 2.2. Several approaches will be tested during the project and their impacts evaluated through the comparison and querying metagenomic software. Indirect deliverables of task 3 are thus implementation in the software of the statistical methodologies.

Risks:

The main risks are related to the very high dimension of the data at hand. Sparse PCA is already well established in the Gaussian (i.e. continuous) setting for a fairly large number of variables, but – to our knowledge – has never been applied to very large dimension. The adaptation to count data will both require methodological extensions (doable within the generalized model framework) and computational adaptations, which may be a major issue so scalability can not be guarantied from now.

Still a pre-selection of the k-mers to be dealt with will always be possible, either by random sampling or based on motif statistics (using R'MES). This will result in a sub-optimal, but still efficient and statistically ground procedure.

Requested resources:

	Operating cost	Justification
INRA	3.0 K€	Missions to Rennes– working meeting, interaction with tasks 2 & 4
	9.0 K€	International congress over the 42 months
	4.0 K€	Publications costs
	4.0 K€	Computer resources (postdoc+ MIGALE platform)
INRIA	2.0 K€	Missions to Paris
	Staff cost	Justification
INRA	93.8 K€	1 Postdoc – 2 years

The person to be recruited will contribute to develop and to implement statistical tools for the classification and the comparison of metagenomic samples. He/she will first have to make him/herself aware of already existing statistical tools, mostly in motif statistics and in the sparse version of the multivariate methods mentioned above. A good statistical background will be required combined with computer programming skills (including low-level languages such as C or C++). Of course a strong interest in bioinformatics and biodiversity is highly desirable.

Task 4: Analysis of metagenomic sequences

Task leader: Olivier Jaillon

Objectives:

This task aims to test, train and validate developments from task 2 and task 3 on real large-scale data, in the present case, data from the Tara Oceans expedition. Two problems related to environmental questions will be more specifically addressed: (1) global survey of plankton diversity and (2) detection of unknown gene families. In addition, this task will evaluate the impact of different sequencing strategies, including emerging ones, in order to provide guidelines for metagenomic projects.

Working program:

Task 4.1: Application to global survey of plankton diversity

Genoscope (Partner 3) is collaborating with marine biologist and oceanographers. With Daniele Ludicone from the Stazione Zoologica Anton Dohrn, Napoli we propose to connect oceanic streams information and genomic data. Thanks to the Tara Oceans project we have in hands genetic information of planktonic organisms sampled at locations chosen according to their oceanographically situations. Several marine regions such as Agulhas rings and Mediterranean Sea circulation flux are already studied in this scope, but we need a better resolution to interpret the results. The measures of genomic similarities between samples generated thanks to methodologies developed in tasks 2 and 3 will be transformed in a geographical map showing similarities between organisms. This map will be analyzed in regards to current knowledge on oceanic streams, and local chemical measures to characterize the distribution of planktonic communities and their interaction with the environment.

Task 4.2: Application to the detection and characterization of families of unknown genes.

We are facing to a recent and increasing number of discoveries of genes having uncharacterized function or no similarities. In the last few years most of these discoveries were in non-classical models especially marine species (Colbourne *et al*. 2011, Chapman *et al*. 2010, Zhang *et al.* 2012). Indeed, in a preliminary but large-scale metagenomic analysis of marine samples we detected a rate of unknown genes ranging from 10 to 80% (Jaillon et al. Submitted). Here, we propose to set up the validation of comparative metagenomic tools and statistical methods of HydroGen in the detection and characterization of families of unknown genes.

We will schedule this process in two steps. First, we will use sequences of unknown genes we already have as representative to interrogate by sequence comparisons (with tools developed in task 2) large collection of environmental sequences that are not annotated. This step will need the statistical evaluation of some similarity measure between (group of) sequences based on the approaches developed in Task 3. Some of the unknown genes can be located in a genome sequence of an identified species; the priority will be given on these genes. We expect to obtain here thousands of families of genes. Members of single families would be related to the representative at different phylogenetic levels, from closely related species (even same species) to less related species. As each family member comes from a sample, it can also be connected to environmental metadata (physical and chemical measures of the environments). We will then obtain a network of information that can be used in a further step to address functional issues about a given gene.

Secondly, we will analyze each family of unknown genes to examine local evolutionary pressures. The information about how a given gene evolves at the nucleotide level – under

purifying selection or not – according to the type of local environment provides important information related to its function. This will be done using confirmed rationale such as PhyloCSF tool (Lin et al. 2011), which was developed in the computational biology group of MIT led by Manolis Kellis (http://compbio.mit.edu/) and with which we will collaborate in this task. This approach has been demonstrated successful when numerous gene sequences from related species can be aligned (Lin et al. 2011)

The PhD student hired (see task 2) will spend several weeks in this group to ensure the successful application of this methodology on environmental data.

Task 4.3. Evaluation of sequencing technologies

Today, Illumina technology is often used to sequence metagenomic samples. This sequencing technique produces very huge amount of short reads that are the primary material of metagenomic analysis. The bioinformatics and statistical tools that will be developed by tasks 2 and 3 mainly target this kind of sequencing, which provides only short fragments of DNA sequences. But depending of the addressed biological question it is not always clear to specify what is the best strategy to produce optimal metagenomic datasets. Based on several environmental samples, of different complexity, this subtask intends to draw general guidelines that could be used as a reference to set up a metagenomic sequencing project. We propose to look at several features such as sequencing depth or library type and to evaluate for each type of datasets the behavior and efficiency of the tools produced by tasks 2 and 3. In particular, statistical metrics and models may need some refinements according to the sequencing features. This work will make them more robust and practical in various contexts.

Furthermore, new sequencing technologies will probably emerge as new standard during the HydroGen project and start to be routinely used. These technologies will provide very long reads, with sometimes no need of an amplification step before sequencing. It is thus extremely important, in the HydroGen project, to take care of the possibilities offered by these new technologies for metagenomic purpose. They won't probably replace high throughput sequencing, but will be used as complementary material. One of the objective of this sub task is thus to perform a technologic watch to anticipate potential impacts in the analysis of metagenomic datasets.

Deliverables:

D4.1 Genomic comparison maps of environmental projects (T0+36)
D4.2 A set of families of gene sequences previously un-described (T0+42)
D4.3 Guideline on sequencing technologies for metagenomic project (T0+24)

Risks:

In case that developments made in HydroGen cannot be applied at large scale, we would have to reduce the scope of the biological objectives. We would have either to use existing tools but that are less efficient, or to analyze less data. In both cases, biological conclusions would be affected. Another theoretical risk is related to the quality and or to the availability of metagenomic data. But concerning the Tara Oceans projects, we have in hands tens of samples sequenced with good quality, and we have funds (OCEANOMICS ANR grants) ensuring a complete sequencing. However, the goal of the HydroGen project is to provide tools and approaches for any environmental genomic project and we are confident that numerous projects will benefit from HydroGen.

Requested resources:

	Operating cost	Justification
CEA	2.0 K€	Missions to Rennes, working meeting, interaction with tasks 2 & 3
	4.5 K	International congress
INRIA	2.0 K€	Missions to Paris
	Staff cost	Justification
CEA	70.0 K€	1 year engineer: Evaluation of sequencing technologies

We request a 1-year engineer to work more precisely on task 4.3. He will have the charge of testing the tools developed inside HydroGen on different metagenomic datasets provided by partner 3 (Genoscope) in order to set up guidelines for metagenomic project.

Task schedule & Deliverables

	Months													
	1-3	4-6	7-9	10-12	13-15	16-18	19-21	22-24	25-27	28-30	31-33	34-36	37-39	40-42
T1														
T2.1														
T2.2														
T3.1														
T3.2														
T3.3														
T4.1														
T4.2														
T4.3														

T1: Management - T2: Bioinformatics tools - T3: Statistical methods - T4: Analysis of metagenomic sequences

Task 1: Management

D1.1	ANR 6 month intermediate reports	T0+6
D1.2	HydroGen web site	T0+6
D1.3	ANR mid-project report	T0+21
D1.4	ANR final report	T0+42

Task 2: Bioinformatics tools

D2.1.1	Metagenomic comparison software – prototype	T0+18
D2.1.2	Metagenomic comparison software – reliable software	T0+30
D2.2.1	Metagenomic querying software – prototype	T0+30
D2.2.2	Metagenomic querying software – reliable software	T0+42

Task 3: Statistics methodology

D3.2	Statistical tools for metagenomic sample clustering	T0+24
D3.1	Statistical tools for reducing high k-mer dimensions	T0+30
D3.3	Statistical tools for metagenomic sample comparison	T0+36

Task 4: Analysis of metagenomic sequences

D4.3	Guideline on sequencing technologies for metagenomic project	T0+24
D4.1	Genomic comparison maps of environmental projects	T0+36
D4.2	A set of families of gene sequences previously un-described	T0+42

3. Strategy of valorization, protection and exploitation of the results.

Scientific valorization

Results of the HydroGen project will be valorized into several complementary ways:

- Scientific publications: strategies and methodologies developed by members of the HydroGen project will be published into the major revues of the domain and partners will attend to the main international congresses for presenting the research issued from the HydroGen project
- Web Site: at the beginning of the project, a web site will be created. Software developed will be made available to the scientific community.

Protection

All software will be deposited at the APP (*Agence de Protection des Programmes*). They will be available under GNU-like licenses. The exact type of license will have to be defined between the partners according to their institute strategies.

Software diffusion

To ensure an efficient promotion of our research work, software that will be released to the scientific community need:

1. To reach a high level of quality
2. To be available for different platforms (Linux, mac-os, windows)
3. To support the feedback of the users

These engineering activities, which are not really the responsibility of a research team and very time-consuming, will be done by subcontracting some of these different tasks to a software company. In our domain, the acceptation of new innovative methods and/or strategies is tightly correlated to the use of robust, user-friendly and multi-platform software. Research impact is thus highly dependent of the quality of the software that is proposed to the community. A solid methodology supported by efficient tools guarantee a better visibility.

Industrial valorization

The INRIA/IRISA GenScale team has a strong partnership with the Korilog Company. In 2013 they have a created a common laboratory (INRIA I-LAB), called KoriScale. Current common research work focuses on intensive sequence comparison with a clear orientation to metagenomic for the next few years. Researches, which will be conducted inside the HydroGen project, have consequently a great potential of valorization, largely exceeding the ocean stream study supported by the HydroGen project. Fundamental statistical and algorithmic tools developed during the project may concern any environmental metagenomic projects.

4. References

Bibliography

S.F. Altschul et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389–3402, 1997

A.M. Cardoso, J. V. Cavalcante, M.E. Cantão, C.E. Thompson, R.B. Flatschart, A. Glogauer, S.N. Scapin, Y.B. Sade, P.I. Beltrão, A.L. Gerber, O.B. Martins, E.S. Garcia, W. De Souza, and A.T. Vasconcelos. Metagenomic analysis of the microbiota from the crop of an invasive snail reveals a rich reservoir of novel genes. PLoS ONE, 7(11):e48505, Nov. 2012

J.A. Chapman, E.F. Kirkness, O. Simakov, S.E. Hampson, T. Mitros, T. Weinmaier, T. Rattei, P.G. Balasubramanian, J. Borman, D. Busam, D, et al. The dynamic genome of Hydra. Nature 464, 592– 596, 2010

Colbourne et al. The Ecoresponsive Genome of *Daphnia pulex*, *Science* 331, 555–561 (2011).

Q. Dai, Y. Yang, T. Wang (2008) Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison. Bioinformatics, 24:2296–2302.

Tom O Delmont, Pascal Simonet, and Timothy M Vogel. Describing microbial communities and performing global comparisons in the 'omic era.6(9):1625–1628, Jun 2012

S.D. Ehrlich and MetaHIT Consortium. Metagenomics of the intestinal microbiota: potential applications. Gastroentérologie Clinique et Biologique – Clinics and Research in Hepatology and Gastroenterology (Sept 2010), Suppl 1 Vol 34, S23-S28.

B.E. Dutilh, R. Schmieder, J. Nulton, B. Felts, P. Salamon, R.A. Edwards, and J.L Mokili. Reference-independent comparative metagenomics usingcross-assembly: crAss.Bioinformatics, Oct 2012

R.C. Edgar. Search and clustering orders of magnitude faster than BLAST, Bioinformatics 26(19), 2460-2461, 2010

D. Fimereli, V. Detours and T. Konopka. TriageTools: tools for partitioning and prioritizing analysis of high-throughput sequencing data. NAR, 41(7):1-8, 2013

K.U. Foerstner, C. Von Mering, S.D. Hooper and P. Bork. Environments shape the nucleotide composition of genomes. EMBO Rep, 6(12):1208–1213, Dec 2005

S. Ishii, M. Yamamoto, M. Kikuchi, K. Oshima, M. Hattori, S. Otsuka, and K. Senoo. Microbial populations responsive to denitrification inducing conditions in rice paddy soil, as revealed by comparative 16S rRNA gene analysis. Applied and Environmental Microbiology, 75 (22) : 7070–7078, Nov. 2009

S. Jaenicke, C. Ander, T. Bekel, R. Bisdorf, M. Dröge, K.H. Gartemann, S. Jünemann, O, Kaiser, L. Krause, F. Tille, M. Zakrzewski, A. Pühler, A. Schlüter, and A. Goesmann. Comparative and joint analysis of two metagenomic datasets from a biogas fermenter obtained by 454-pyrosequencing. PLoS ONE, 6(1):e14519, Jan2011

O. Jaillon, C. Chica, E.M. Novoa, P. Hingamp, E. Pelletier, J. Poulain, J.M. Aury, B. Noel, S. Audic, M. Bansal, P. Bento, M. Carmichael, C. Da Silva, J. Decelle, C. Dimier, I. Ferrera, M. Kantinka, K. Labadie, S. Romac, G. Samson, S. Acinas, F. Benzoni, P. Bork, C. Bowler, M. Follows, G. Gorsky, N. Grimsley, D. Iudicone, S. Kandels-Lewis, F. Not, S. Pesant, J. Raes, C. Sardet, Sabrina S., L.

Stemmann, M. Sullivan, S. Sunagawa, E. Karsenti, M. Kellis, M. Sieracki, H. Ogata, C. de Vargas, P. Wincker , Unknown genes from marine plankton are mainly eukaryotic, Submitted to Nature Communication, may 2014

W. J Kent. BLAT : the BLAST-Like alignment tool. Genome Research,12(4):656–664, Mar 2002

M.F. Lin, P. Kheradpour, S. Washietl, B.J. Parke, J.S. Pedersen, M. Kellis, Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. Genome Research, 21(11):1916-28, Nov. 2011

M.F. Lin, I. Jungreis, M. Kellis, PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics, 1;27(13):i275-82, Jul. 2011.

N. Maillet, C. Lemaitre, R. Chikhi, D. Lavenier and P. Peterlongo. Compareads: comparing huge metagenomic experiments. BMC Bioinformatics, 13(19):S10, 2012

S. Mirete, CG. De Figueras, and JE. Gonzalez-Pastor. Novel nickel resistance genes from the rhizosphere metagenome of plants adapted to acid mine drainage. Applied and Environmental Microbiology, 73 (19):6001–6011, Oct. 2007

VH. Nguyen, D. Lavenier. PLAST: parallel local alignment search tool for database comparison, BMC Bioinformatics, vol 10, no 329, 2009

J.A. Port, J.C. Wallace, W.C. Griffith and E. M Faustman. Metagenomic profiling of microbial composition and antibiotic resistance determinants in puget sound. PLoS ONE, 7(10):e48000, Oct 2012

J. Raes, J.O. Korbel, M.J. Lercher, C. Von Mering and P. Bork. Prediction of effective genome size in metagenomic samples. Genome biology, 8(1):R10, Jan. 2007

G. Reinert, S. Schbath, M. Waterman (2005). *Applied Combinatorics on Words*. volume 105 of *Encyclopedia of Mathematics and its Applications*, chapter Statistics on Words with Applications to Biological Sequences. Cambridge University Press.

S. Robin, F. Rodolphe, S. Schbath, (2005) DNA, words and models. Cambridge university Press.

S. Schbath, M. Hoebeke, (2011). *Advances in genomic sequence analysis and pattern discovery*. (L. Elnitski, O. Piontkivska, and L. Welch, ed.), chapter R'MES: a tool to find motifs with a significantly unexpected frequency in biological sequences. Science, Engineering, and Biology Informatics, vol. 7. World Scientific.

M. Shakya, C. Quince, J.H. Campbell, Z.K. Yang, C.W. Schadt, and M. Podar. Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. Environmental Microbiology, Feb. 2013

K. Song, J. Ren, Z. Zhai, X. Liu, M. Deng, F. Sun (2013) Alignment-free sequence comparison based on next-generation sequencing reads. J. Comput biol, 20(2):64-79.

J.C. Venter-1 et al., The sorcerer II Global Ocean Sampling Expedition: Northwest atlantic through eastern tropical pacific. Plos Biol,5(3):e77, Jan. 2007

J.C. Venter-2 et al., The sorcerer II Global Ocean Sampling Expedition: Expanding the universe of protein families. Plos Biol, 5(3):e16, Jan. 2007

Y. Wang, L. Liu, L. Chen, T. Chen, F. Sun (2014) Comparison of Metatranscriptomic Samples Based on *k*-Tuple Frequencies. PLoS ONE 9(1): e84348

JC. Wooley, A. Godzik, and I. Friedberg. A primer on metagenomics. PLoS Comput Biol, 6 (2):e1000667, 2010

G. Zhang et al. The oyster genome reveals stress adaptation and complexity of shell formation. Nature ;490(7418):49-54, Oct. 2012

H. Zou, T. Hastie, R. Tibshirani. Sparse principal component analysis. Journal of computational and graphical statistics, 15(2), 265-286, 2006

Publications INRIA / IRISA / GenScale (5)

G. Rizk, D. Lavenier, T. Chikhi, DSK: k-mer counting with very low memory usage Bioinformatics, Volume 29, Issue 5, 2013

K. R. Bradnam, J. N. Fass et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species, GigaScience, vol.2, 2013

N. Maillet, C. Lemaitre, R. Chikhi, D. Lavenier, P. Peterlongo, Compareads: comparing huge metagenomic experiments BMC Bioinformatics 2012, 13 (Suppl 19):S10

R. Chikhi, G. Chapuis, D. Lavenier, Parallel and memory-efficient reads indexing for genome assembly, Parallel Processing and Applied Mathematics, LNCS Vol. 7204, 2012, pp 272-280

P. Peterlongo, R. Chikhi. Mapsembler, targeted and micro assembly of large NGS datasets on a desktop computer, *BMC Bioinformatics*, 2012, 13 (1), pp. 48.

Publications INRA / MIG / MIA AgroParisTech (5)

Robin, S., Rodolphe, F. and Schbath, S. (2005) DNA, words and models. Cambridge University Press.

Schbath, S. and Hoebeke, M. (2011). Advances in genomic sequence analysis and pattern discovery. (L. Elnitski, O. Piontkivska, and L. Welch, ed.), chapter R'MES: a tool to find motifs with a significantly unexpected frequency in biological sequences. Science, Engineering, and Biology Informatics, vol. 7. World Scientific.

Reinert, G., Schbath, S. and Waterman, M. (2005). Applied Combinatorics on Words. volume 105 of Encyclopedia of Mathematics and its Applications, chapter Statistics on Words with Applications to Biological Sequences. Cambridge University Press.

Chiquet, J., Mary-Huard, T. and Robin, S. (2014) Structured Regularization for conditional Gaussian Graphical Models, arXiv:1403.6168

J Chiquet, Y Grandvalet, C Ambroise (2011) Inferring multiple graphical structures, Statistics and Computing 21 (4), 537-553.

Publications CEA / Genoscope (5)

Hingamp P, Grimsley N, Acinas SG, Clerissi C, Subirana L, Poulain J, Ferrera I, Sarmento H, Villar E, Lima-Mendez G, Faust K, Sunagawa S, Claverie JM, Moreau H, Desdevises Y, Bork P, Raes J, de Vargas C, Karsenti E, Kandels-Lewis S, Jaillon O, Not F, Pesant S, Wincker P, Ogata H. Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. ISME J. 2013

Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, Sullivan M, Arendt D, Benzoni F, Claverie JM, Follows M, Gorsky G, Hingamp P, Iudicone D, Jaillon O, Kandels-Lewis S, Krzic U, Not F, Ogata H, Pesant S, Reynaud EG, Sardet C, Sieracki ME, Speich S, Velayoudon D, Weissenbach J, Wincker P; Tara Oceans Consortium. A holistic approach to marine eco-systems biology. PLoS Biol. 2011 Oct;9(10):e1001177.

Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H, Hingamp P, Ogata H, de Vargas C, Lima-Mendez G, Raes J, Poulain J, Jaillon O, Wincker P, Kandels-Lewis S, Karsenti E, Bork P, Acinas SG. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. Environ Microbiol. 2013 Aug 18.

Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, Bento P, Da Silva C, Labadie K, Alberti A, Aury JM, Louis A, Dehais P, Bardou P, Montfort J, Klopp C, Cabau C, Gaspin C, Thorgaard GH, Boussaha M, Quillet E, Guyomard R, Galiana D, Bobe J, Volff JN, Genêt C, Wincker P, Jaillon O, Roest Crollius H, Guiguen Y. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun. 2014 Apr 22;5:3657.

Flot JF, Hespeels B, Li X, Noel B, Arkhipova I, Danchin EG, Hejnol A, Henrissat B, Koszul R, Aury JM, Barbe V, Barthélémy RM, Bast J, Bazykin GA, Chabrol O, Couloux A, Da Rocha M, Da Silva C, Gladyshev E, Gouret P, Hallatschek O, Hecox-Lea B, Labadie K, Lejeune B, Piskurek O, Poulain J, Rodriguez F, Ryan JF, Vakhrusheva OA, Wajnberg E, Wirth B, Yushenova I, Kellis M, Kondrashov AS, Mark Welch DB, Pontarotti P, Weissenbach J, Wincker P, Jaillon O, Van Doninck K. Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga. Nature. 2013 Aug 22;500(7463):453-7.