A geometric view of Biodiversity: scaling to metagenomics

Pierre Blanchard, Philippe Chaumeil, Jean-Marc Frigerio, Frédéric Rimet, Franck Salin, Sylvie Thérond, Olivier Coulaud, Alain Franc

A geometric view of Biodiversity: scaling to metagenomics

Pierre Blanchard, Philippe Chaumeil, Jean-Marc Frigerio, Frédéric Rimet, Franck Salin, Sylvie Thérond, Olivier Coulaud, Alain Franc

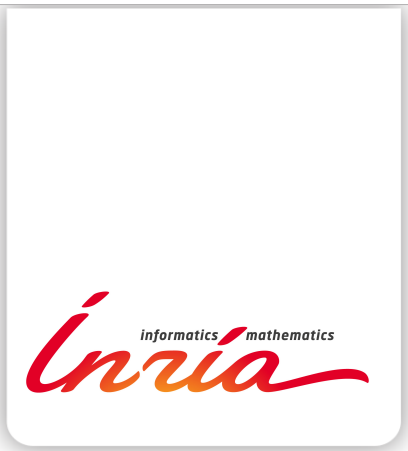
RESEARCH REPORT

N° 9144

January 2018

Project-Teams Pleiade & HiePACS

ISSN 0249-6399 ISRN INRIA/RR--9144--FR+ENG



A geometric view of Biodiversity: scaling to metagenomics

Pierre Blanchard*†, Philippe Chaumeil*†, Jean-Marc Frigerio*†, Frédéric Rimet‡, Franck Salin*†, Sylvie Thérond§, Olivier Coulaud¶, Alain Franc*†‖

Project-Teams Pleiade & HiePACS

Research Report n° 9144 — January 2018 — 16 pages

Abstract: We have designed a new efficient dimensionality reduction algorithm in order to investigate new ways of accurately characterizing the biodiversity, namely from a geometric point of view, scaling with large environmental sets produced by NGS ($\sim 10^5$ sequences). The approach is based on Multidimensional Scaling (MDS) that allows for mapping items on a set of n points into a low dimensional euclidean space given the set of pairwise distances. We compute all pairwise distances between reads in a given sample, run MDS on the distance matrix, and analyze the projection on first axis, by visualization tools. We have circumvented the quadratic complexity of computing pairwise distances by implementing it on a hyperparallel computer (Turing, a Blue Gene Q), and the cubic complexity of the spectral decomposition by implementing a dense random projection based algorithm. We have applied this data analysis scheme on a set of 10^5 reads, which are amplicons of a diatom environmental sample from Lake Geneva. Analyzing the shape of the point cloud paves the way for a geometric analysis of biodiversity, and for accurately building OTUs (Operational Taxonomic Units), when the data set is too large for implementing unsupervised, hierarchical, high-dimensional clustering.

Key-words: Biodiversity, metabarcoding, Multidimensional Scaling, Singular Value Decomposition, Random Projection

* BIOGECO, INRA, Univ. Bordeaux, 33610 Cestas, France
† Pleiade team - INRIA Bordeaux-Sud-Ouest, France
‡ UMR Carrtel, INRA, Thonon-les-Bains, France
§ IDRIS, CNRS, Orsay, france
¶ HiePACS team, Inria Bordeaux-Sud-Ouest, France
‖ Corresponding author, alain.franc@inria.fr

RESEARCH CENTRE
BORDEAUX – SUD-OUEST

200 avenue de la Vieille Tour
33405 Talence Cedex

Caractérisation géométrique de la biodiversité : passage à l'échelle en métagénomique

Résumé : Nous avons conçu un algorithme de réduction de la dimension pour explorer de nouvelles voies pour une caractérisation précise de la biodiversité, ici par une approche géométrique, qui satisfait aux critères de passage à l'échelle pour les jeux de données produits par NGS (actuellement 10^5 reads). Cette approche est basée sur la technique dite "Multidimensional Scaling", qui permet de projeter les éléments à étudier sur un ensemble de n points dans un espace euclidien de faible dimension, connaissant leurs distances respectives. Nous avons calculé toutes les distances deux à deux entre reads d'un échantillon environnemental, réalisé une MDS du tableau de distances, et analysé les projections sur les premiers axes par des techniques de visualisation. Nous avons abordé la question de la complexité quadratique du calcul des distances deux à deux en réalisant les calculs dans un Centre National disposant d'une machine hyperparallèle (Turing, une IBM Blue Gene Q), et la complexité cubique de la décomposition spectrale dans la MDS en utilisant un algorithme de projection aléatoire dense. Nous avons appliqué cette procédure à un jeu de 10^5 reads d'un échantillon environnemental de diatomées du lac Léman. L'analyse de la forme du nuage de points obtenu ouvre la voie vers une analyse géométrique de la biodiversité, et une construction rigoureuse d'OTUs (Operational Taxonomic Units) lorsque le jeu de données est trop grand pour mettre en oeuvre les méthodes de classification ascendante hiérarchique, non supervisée.

Mots-clés : Biodiversité, Métabarcoding, Multidimensional Scaling, Décomposition en Valeurs Singulières, Projection aléatoire

Contents

Deciphering the diversity of life has been a longstanding guiding thread in biology [May82]. Diversity relies on differences and dissimilarities. A community is the set of organisms sharing a common habitat [RM00]. There has been numerous definitions of the diversity of a community, excellently reviewed in [Mar03]. Biodiversity is not random, but organized in a given number of patterns [Hey95]. Identifying those patterns shaped by evolutionary forces is one of the key challenges in biology [Lev92, Gas00]. All these approaches rely on the simplification of the whole diversity as an index, be it the Shannon index, the Simpson index, or even the number of different species. A first objective of this work is to characterize diversity as the shape of a point cloud, where a point is an organism in a community, and localized at a distance from neighbors which represents their dissimilarities. Such an approach is called here the *geometric view on biodiversity* because a unique index or a set of indices is replaced by the shape of a point cloud. Rigorously defining the shape of a point cloud is not an easy task, intertwining computer vision and machine learning [Sze11]. Here, we address a simplification of this question, focusing on one type of shapes which are classical in numerical taxonomy [SS73]: organized sets of clusters. A second objective of this paper is to propose a connection between High Performance Computing and molecular based numerical taxonomy, in order to handle with highest accuracy all the information provided by very large datasets produced by NGS facilities. This connection is established by both using massive parallelization and providing efficient algorithms for linear algebra behind dimensionality reduction of massive data.

1 Introduction

Molecular based taxonomy: For a couple of centuries, the diversity of life has been organized as taxonomic systems, based on nested clustering [SS73]. This relied mainly as an expertise on key dissimilarities between some phenotypic traits, and systematics was a cornerstone of natural history. More recently, systematics has shifted towards molecular systematics [HMM96], where distances between sequences have been organized along phylogenies, which model the history of speciation in a given group [Fel04]. Slightly later, so called barcoding has been an international effort to normalize sequencing (primers choice) to encompass as much as possible of the diversity of all organisms, using very large and shared databases [HCBd03, HFSa09, PAAe12]. Such an approach is perfectly suited for the geometric view of biodiversity as a replicable procedure, because there exists algorithms for computing distances between two sequences. In this work we focus on a molecular based construction which scales with massive data sets produced by NGS. Diversity characterization is obtained by visualization of the produced point clouds by projection in low dimension spaces, as finding a shape in a high dimensional space still remains an open question.

Metabarcoding: As soon as molecular diversity has been put in relation with taxonomic diversity, several studies have shown that most of the existing diversity had been hidden to our eyes. This is particularly true for microbial diversity in the oceans [LGRVPAM01, SMH$^+$06]. In parallel, metagenomics has emerged as a technique to sequence simultaneously all genomes in a given microbial community. Such a technique has been developed as metabarcoding [HSZ$^+$11, BPC$^+$12, TCHR12, KFR$^+$14, DBC$^+$14, JDA$^+$14, PLE14] by amplifying and sequencing all the sequences of a given marker of taxonomic interest in a given community. Here, we develop some data analysis on metabarcoding of environmental samples of freshwater diatoms. We have selected this clade as it is one of the clades of microbial organisms for which an expertise exists for optical based taxonomic identification [Man99]. Hence, consistency between morphological based and molecular based taxonomy can be evaluated.

2 Data analysis

As in machine learning, two main families of methods exist to assign taxonomically a set of sequences:

- supervised learning, where an unknown sequence is assigned to a taxon by comparison with a reference database, where sequences have been taxonomically annotated
- unsupervised learning, where the organization of the diversity itself is sought for, as in clustering. This is traditionally called OTU picking in metabarcoding studies [BPC$^+$12].

There exists classical and efficient tools for each of this task. BLAST [AGM$^+$90] is by far the most used tool for supervised learning. However, there is no equivalent golden standard for unsupervised clustering. It should be nested agglomerating clustering, which suffers from computation load as computing all distances between n sequences and performing agglomerative clustering are each quadratic with n, i.e. the complexity is in $\mathcal{O}(n^2)$ [Mü13]. Hence, many heuristics have been used. Among them, one of the mostly used is an adaptation of k-means [Edg10]. It can handle huge datasets, as it can work in streaming, but it suffers from some flaws, as the fact that the result depends on the order the sequences have been presented. Swarm [MRQ$^+$14] has solved some of these issues, but still relies on some heuristics. However, there is no better tool for huge data sets (millions of sequences) for clustering. Here, we compute exact distances between

sequences, and run Multidimensional Scaling as a way to visualize the structure of the diversity as given by pairwise distances.

Historically, Multidimensional Scaling has been derived to find configurations of points in Euclidean space of small dimension the shape of which fits as much as possible a set of pairwise distances [Tor52]. However, the machine learning algorithms used for building a point cloud knowing pairwise distances, as well as just computing the distances between all pairs of reads, have a complexity respectively cubic and quadratic to the number of specimens. Such a complexity has been acknowledged as a difficulty for dealing with full matrices of pairwise distances, and several procedures have been designed to circumvent it (see [AK13] and references therein). In this work, we implement an algorithm to circumvent the complexity of the problem by computing pairwise distances on an hyperparallel supercomputer, and by calculating the desired eigenpairs by methods based on random projection [Vem04, HMT11].

2.1 A generic approach for large biological datasets

The pipeline we have designed and implemented is the sequence of following steps:

- compute pairwise distances between all pairs of n reads,
- run Multidimensional Scaling on the distance matrix,
- visualize the first r components with $r \ll n$.

The novelty lies in an implementation on large data sets, *i.e.*, scaling up to data sets produced by NGS sequencing operations, currently $n \simeq 10^5$ reads. Computing pairwise distances between reads is quadratic in n, and requires $\mathcal{O}(Kn^2)$ elementary operations, K being a constant quadratic in the length of the reads. Running MDS requires computing r eigenpairs of a $n \times n$ matrix, which can be done in $\mathcal{O}(rn^2)$ operations. We address this complexity issue by

- parallelization of the computation of pairwise distances on an hyperparallel machine (an IBM Blue Gene Q)
- using a MDS approach powered by random projection and efficient C++ implementation.

The time needed for each of these steps is given in table 1.

Sample (Lake Geneva)	compute distances S-W on *Turing*	transfer and convert *iRods→plafrim*	subsample 10^4 reads	perform MDS full SVD
10^5 reads	4h	20min + 1h40	30s	20min

Table 1: Some average running times related to expensive operations involved in the operational chain.

2.2 Scalable pairwise distances computation

There are several ways to compare sequences (see e. g. [Gus97, Yan06]). The dissimilarity used here is computed from a local alignment score [Gus97, p. 232] using Smith-Waterman algorithm [SW81]. If the length of the sequences is p, the algorithm scales like $\mathcal{O}(p^2)$. As it is quadratic in the number n of sequences, computing the whole matrix has a complexity of $\mathcal{O}(n^2p^2)$. We have written a program in C, called `disseq`, which takes as inputs two fasta files, of length m and n each, and returns the $m \times n$ matrix $D = [d_{ij}]$ where d_{ij} is the distance between sequence i

of first file and j of second file. This program has been parallelized with MPI (Message Passing Interface) as a map-reduce process, in a program called `mpi-disseq`, as all distances can be computed independently. `mpi-disseq` has been run on a BlueGene Q (IBM) located at IDRIS to compute the matrix D of all pairwise distances ($\sim 5.10^9$ values). Its architecture is made of 6 racks, of 1,024 nodes each, with 16 cores per node, hence 98,304 cores. We have used Turing with $2^{14} = 16,384$ cores (one rack). Turing peak power is 1,258 Pflops/s. One advantage of such a choice, beside massive parallelization, is a low energy consumption. Such an architecture is particularly suitable for massive embarrassingly parallel jobs. The program has been tested on Avakas (Mésocentre de Calcul Intensif Aquitain, 264 computing nodes, 12 cores per node), and then ported to Turing. It scales perfectly.

3 A random projection-based spectral decomposition

In this section, D denotes the pairwise distance matrix and G the Gram matrix built from it (see section 6). Next step is to perform MDS on D. The calculation is presented in section Material and Methods. It amounts to perform a SVD of a Gram matrix G built from the distance matrix D, and of same dimension. In this section we describe an efficient approach for computing the spectral decomposition of the real symmetric matrix G at a quadratic cost in n. We put the emphasize on the genericity of the approach and its competitive numerical performance (see [Bla17] for details).

3.1 Standard iterative approach

The standard methods used to compute the full eigenvalue decomposition (EVD) of an arbitrary n-by-n real matrix usually take $\mathcal{O}(n^3)$ operations. This complexity makes those methods computationally untractable to large data sets. However, there exists well-known iterative techniques that compute the first r eigenpairs at a $\mathcal{O}(n^2 r)$ cost, *e.g.*, Arnoldi or Lanczos algorithm [Sor97]. Despite presenting several flaws, these variants provide an exact spectral decomposition.

3.2 A *random projection*-based algorithm

Our novel contribution to MDS consists in using low-rank approximation techniques based on random projection, to compute an approximate rank-r spectral decomposition of G. Random projection, along with random sampling, belongs to a broader class of dimensionality reduction tools known as random sketching [Woo14]. Although most fast approaches to MDS rely on random sampling [Pla05], to our knowledge there are very few contributions to this field that involve random projection if any. Despite involving more intensive computations than random sampling, random projection presents many benefits such as better accuracy, robustness and low variability. Low-rank approximation techniques based on random projection were made popular by a SIAM review paper by Halko et al. [HMT11]. Among theses algorithms, the randomized Singular Value Decomposition (or randomized SVD) computes an approximate rank-r SVD of a m-by-n matrix in $\mathcal{O}(mnr)$ operations. This algorithm can be used to compute the first eigenpairs of G, by simply remembering that for a symmetric real matrix singular values σ can be related to the eigenvalues by $\sigma = |\lambda|$.

3.3 Numerical performance

This approach represents a significant alternative to standard EVD algorithms, as it shares the same complexity, namely $\mathcal{O}(rn^2)$, but usually performs significantly better and parallelizes

straightforwardly. However, the input matrix has to fulfill a few criteria for the method to work properly. G should be low-rank, *i.e.*, $r \ll n$, and have a fast decreasing spectrum or a low stable-rank, *i.e.*, $st(G) = \|G\|_F^2/\|G\|_S^2 \ll n$, *i.e.* a low ratio between the squared Frobenius norm $\|G\|_F^2 = \sum_{i\leq n} \sigma_i^2$ and the squared spectral norm $\|G\|_S^2 = \max_{i\leq n} \sigma_i^2$. Finally, as stated in [HMT11], even though the algorithm has a non-zero chance to fail, tight Frobenius and spectral error bounds were shown to hold with high probability, if G verifies the aforementioned conditions.
-

4 Visualization of a large diatoms sample

The main goal of our approach is to achieve an efficient and handy visualization technique that provides relevant information on the diversity of some real-life samples, currently of about 10^5 sequences. Due to the longstanding coevolution between clustering and numerical taxonomy, a possible analysis of the shape of the point cloud consists in identifying clusters. Clustering is an immense domain, and we will not enter here into the discussion for the best methods knowing a pairwise distance matrix: this is ongoing work. We focus in this report on studying the concentration of the reads in some subsets in the projection on the first few dimensions of MDS.

4.1 Dimensionality reduction

When the dimension of a space is large, the so called *curse of dimensionality* enters into the game. In brief, it can be summarized by telling that there is much room in a high-dimensional space. For example, two opposite vertices of an hypercube are at distance $\sqrt{n}$ (i.e. the distance between two points in a symmetric body of unit volume can be arbitrarily large), whereas the volume of the hypersphere of radius one shrinks to zero: the probability to have a neighbor at distance one or less is negligible [Ize08, Wan12]. This can impact any algorithm based on distances. Therefore, a first step is often to reduce the dimension of the problem, by building a mapping of the original point cloud on a Euclidean space of far lower dimension. This is called dimensionality reduction, and is a key step in machine learning [Ize08, LV07, Mur12]. MDS is the tool for visualizing the geometry of the point cloud associated to distances between reads in low dimension. Hence we present such a visualization by projection on first axis of MDS, and provide some hints for interpretation by a comparison with supervised clustering. We do not report here on unsupervised clustering, which is deferred to further work. We have worked with a full sample named L_6 ($n = 99,594$, see sample description in Materials and Methods section). We have built the associated point cloud by mean of a rank $r = 50$ randomized SVD. Here, we show representations of the full sample in 3D for easy visualization of the point cloud. In such a representation, each point represents a read. We have used a supervised clustering technique (a program called `diagno-syst`, see [FRB+16]) to assess a taxonomical name to each read for which it has been possible, to test whether the clusters were related to species. Then, we discuss the visualization of the cloud in higher dimensions.

4.2 Reads concentration in low dimension

Because of the number of points, several reads are projected on the same pixel in regions with high concentration of reads in first axis. We have quantified this by hexagonal binning: the plane formed by axis i and j (with $(i,j) \in (1,2);(1,3);(2,3)$) is tessalated by hexagons. Then, the number of points per hexagon is counted, and these figures are displayed, in logarithmic scale. This is presented in figure 1. In particular, this figure shows that only very specific regions of

the full domain contain most of the individuals and that they are surrounded by regions of far lower density. The pattern is an archipelago of islands of high density surrounded by an ocean of low density. It may be tempting to associate high density islands to clusters and clusters to species.

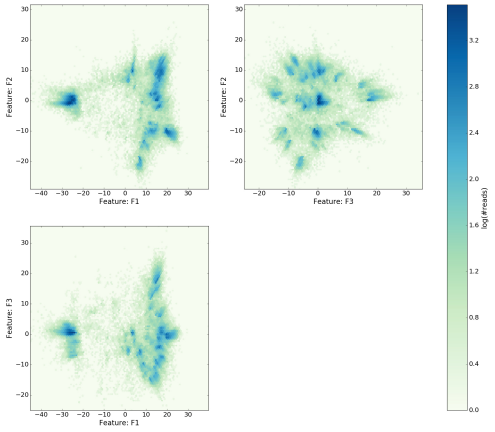


Figure 1: Concentration of the population of a full sample represented in log-scale on the first thee axis.

4.3 Interpretation with supervised classification

In order to validate our approach, and test whether high density islands correspond to reads belonging to a same species, with ideally a one to one correspondence between islands and species, we have colored the points according to an information provided by supervised clustering on all the reads. Indeed, we have at disposal a reference database for diatoms, which is fairly good for Geneva lake [RCK+16]. We have computed all pairwise distances between the reads in our sample ($\simeq 10^5$), and the taxonomically annotated reads in the reference database ($\simeq 2.10^3$). This is intensive computing too, which has been executed on Turing, with `MPI-disseq` as well. Then, we have selected an homology gap (classically 97%), and, for each read in our sample, selected all reads in the reference database which were at distance lower than the homology gap. If they all belong to a same species, the read has been annotated with this species name. If not, it has been annotated as ambiguous status. This has been done with a program called `diagno-syst`, as a work available in [FRB+16]. Then, we have colored the individuals *w.r.t.* their species as identified by supervised clustering in different colors for different species. (see figure 2). We have checked as well that the species identified this way in the community have been identified optically too. This supports a consistency between morphological based and molecular based taxonomy.

4.4 Representation in many dimensions

In order to visualize the point cloud in more dimensions than 3D, a common alternative to 2D plots is the representation in *parallel coordinates* such as the one displayed on Figure 3. This technique offers various advantages as it decouples the dimensions and displays them on a single axis. Due to significant density of the cloud, we make use of advanced rendering techniques (opacity, brushing and bundling) provided by the javascript libraries `d3.js` and `paracoords.js`

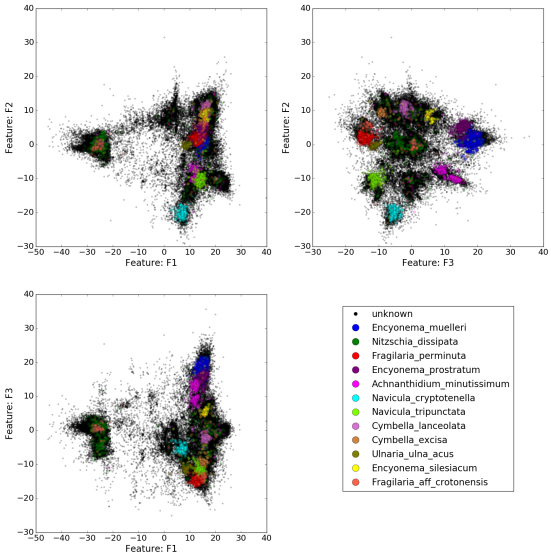


Figure 2: Concentration of the population of a full sample represented in log-scale on the 3 first pairs of features.

in order to better represent the concentration of reads. The clustering of the reads can again be confirmed by observing that individuals of the same species follow a similar path along the first few dimensions, while spreading only on a fraction of the full domain. This observation is made clearer by the observation of each species separately. The consistency of clustering as seen through bundles of trajectories of all reads belonging to the same species with species delineation is shown here up to the first ten dimensions, which permits to envisage such an approach for pattern discovery related to OTU beyond visual inspection in the first three axis.

5 Conclusions and perspectives

We have implemented Multidimensional Scaling of the matrix which contained exact distances between all pairs or reads of an environmental sample of $\simeq 10^5$ reads. This has permitted to visualize the points cloud associated to those distances in low dimension spaces. We have shown that points were clustered in islands of high density regions forming an archipelago within an ocean of low density regions. We have shown by supervised classification that some clusters could be associated to those species which were optically identified in the sample and in the reference database. This establishes consistency between supervised and unsupervised approaches, and paves the way for using unsupervised methods when taxonomic information for supervised approaches is lacking, which will greatly expand the scope of these methods.

It can be observed as well that some islands could not be assigned to species by supervised clustering. It may be tempting to hypothesize that those clusters correspond to species which are present in the sample, but not in the reference database. More precisely, some parts of the cloud contains unknown individuals but still exhibits a certain structure and a significant density, which suggests that a deeper analysis of the cloud is possible. Supervised method permit a dictionary between molecular based patterns as derived in metabarcoding and previous

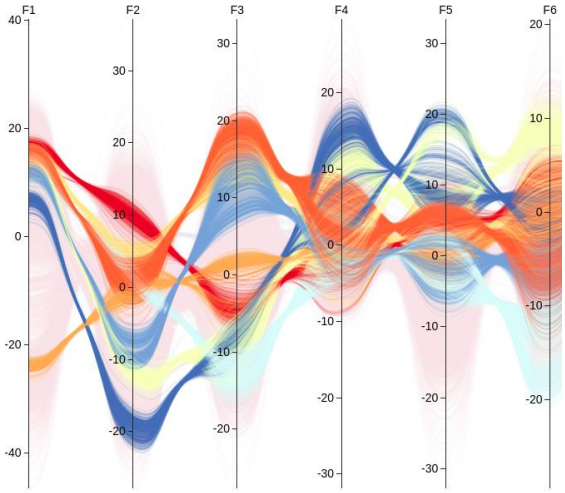


Figure 3: Representation of the predominant identified species within the full sample using parallel coordinates over the first 6 features produced by random projection-aided MDS. Legend: Encyonema muelleri (orange), Fragilaria perminuta (light blue), Nitzschia dissipita (dark yellow), Encyonema prostratum (yellow), Achnanthidium minutissimum (blue), Navicula cryptotenella (dark blue), Navicula tripunctata (light yellow), Cymbella lanceolata (red). The remaining reads appear in light red in the background.

patterns in biodiversity studies as derived in Natural History. However, one of the most critical shortcoming of supervised methods is that reference databases will never be complete, because of the tremendous number of species, most of them still unknown. Hence, there is an urgent need for unsupervised methods, which are consistent with supervised methods when taxonomic information is available. We have proposed here such a method by an analysis of high density regions in projection in a low dimension space. Nest step may be the use of image analysis techniques to identify and characterize those high-density islands.

Classically, unsupervised clustering of reads for metabarcoding is done by greedy methods, mentioned in the introduction of section 2 (a couple of them are included as choices in Mothur [SWR$^+$09] or Qiime [CKS$^+$10]). Once again, those methods rely on heuristics, and their accuracy are not assessed for very large datasets. Many different exact methods can be implemented provided one accepts to invest in computation load for a better accuracy. One can mention a few: nested aggregative clustering [Mü13], selecting a threshold θ and build a graph $G = (V, E)$ with V being the set of reads, and $(i, j) \in E \Leftrightarrow d(i, j) \leq \theta$, and build the connected components of G (this is linear in time with n), or more elaborate methods on the same graph like spectral clustering to derive communities [vL07, GN02].

Next steps are not only to compare, but to associate all these methods and tools, in order to build Operational taxonomic Unit with best accuracy on very large data sets, as consistent as possible with our knowledge of the diversity of life as we have inherited it from centuries of studies in Natural History, and as automatized and sound as possible for exploring the unknown diversity.

6 Material and methods

6.1 A sample from Lake Geneva

We have considered 10 environmental samples, denoted L_t, that were collected from Lake Geneva at about monthly intervals at times $t = 1, \ldots, 10$ between April 2012 and March 2013 in order to investigate a seasonal dynamics. Amplicons of chloroplastic marker rbcL have been produced for each sample by DNA extraction, amplification, and sequencing on a Ion Torrent PGM (see [KFR$^+$13] for protocols). The various samples contain from about 7×10^4 to 1.4×10^5 reads. The diversity of each sample has been assessed both optically, and by supervised clustering of reads by mapping on a dedicated reference database (see [RCK$^+$16] for the reference database, and [FRB$^+$16] for the algorithm for mapping reads, and its implementation.) A fraction only of diversity can be assessed with these classical tools.

6.2 Multidimensional Scaling

The method implemented here is *classical multidimensional scaling*[1] (see [BG05, Ize08] for a recent survey, and [CC01] for a seminal monograph) which follows Torgerson's work [Tor52]. Let us have a set of n items with $i \in V = \{1, n\}$, and a distance d_{ij} between items i and j. Then, (V, d) is a finite metric space. For a given dimension $r \in \mathbb{N}$, classical MDS is finding a map $x : i \longrightarrow x_i \in \mathbb{R}^r$ such that $\|x_i - x_j\|$ is as close as possible to d_{ij}. If the distances d_{ij} are such that the corresponding Gram matrix G is definite positive, the solution is well known [CC01], and is implemented in three steps

1. The scalar product $\langle x_i, x_j \rangle$ can be computed from the distances only, as

$$\langle x_i, x_j \rangle = -\frac{1}{2}\left(d_{ij}^2 - \frac{1}{n}\sum_i d_{ij}^2 - \frac{1}{n}\sum_j d_{ij}^2 + \frac{1}{n^2}\sum_{i,j} d_{ij}^2 \right)$$

   The scalar products $\langle x_i, x_j \rangle$ are the elements of the Gram matrix $G = [\gamma_{ij}]$ with $\gamma_{ij} = \langle x_i, x_j \rangle$

2. compute the eigenvalues and eigenvectors of Gram matrix G

$$Gu_\alpha = \lambda_\alpha u_\alpha, \qquad \lambda_1 \geq \ldots \geq \lambda_n \geq 0$$

   (As G is a Gram matrix, all its eigenvalues are non negative, as there is a $n \times n$ matrix X such that $G = XX'$)

3. Let us denote $\Sigma = \Lambda^{1/2}$ where Λ is the diagonal matrix with $(\lambda_\alpha)_\alpha$ on its diagonal. Then compute

$$X = U\Sigma$$

Then, the best representaton of D as distances within a point cloud in $\mathbb{R}^k$ embedded with standard inner product is X_r, which is the extraction of the first r columns of X. x_i is the $i-$th row of X_r. The best rank k approximation of G is $G_r = X_r X_r'$.

[1] There is another procedure bearing the same name, called nonmetric MDS due to Kruskal [Kru64], based on an optimization scheme, which currently cannot be implemented for more than a few thousands items.

6.3 Observations

Two observations can be made:

- In general, not all eigenvalues of G are non negative. If all eigenvalues are non negative, then (V, d) can be embedded isometrically in a Euclidean space, and a best low dimensional approximation can be computed as the Principal Component Analysis of the point cloud [MKB79]. The condition can be seen directly on the distance matrix, and involves signs of Cayley-Menger determinents (see [LLMM14, thrm 2.1, p. 15]).
- As far as calculation is concerned, step 2 is the most demanding: computing eigenpairs of the Gram matrix. We have implemented here a procedure relying on random projection (see [Vem04] for a survey).

Acknowledgments: Computation at IDRIS has been supported by DARI project *Biodiversiton* `i2015037360`. Experiments presented in this paper were carried out using the PlaFRIM experimental testbed, being developed under the Inria PlaFRIM development action with support from LABRI and IMB and other entities: Conseil Régional d'Aquitaine, FeDER, Université de Bordeaux and CNRS (see https://plafrim.bordeaux.inria.fr/). Sequencing has been performed at the Genome Transcriptome Facility of Bordeaux (grants from the Conseil Régional d'Aquitaine n°20030304002FA and 20040305003FA, from the European Union FEDER n°2003227 and from Investissements d'Avenir ANR-10-EQPX-16-01) and with support of ONEMA project on Mayotte. Humid lab (DNA extraction, PCR) have been done at UMR Carrtel, under guidance of Agnès Bouchez, with support of ONEMA project Mayotte. We acknowledge an *Investissement d'Avenir* grant of the Agence Nationale de la Recherche (CEBA: ANR-10-LABX-25-01). We acknowledge the nice atmosphere of network R-Syst, and support by INRA divisions EFPA and SPE, for methodological developments on metabarcoding, as well as its database infrastructure.

References

[AGM+90] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[AK13] Y. Aflalo and R. Kimmel. Spectral Multidimensional Scaling. *PNAS*, 110(45):18052–18057, 2013.

[BG05] I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling*. Springer Series in Statistics. Springer, second edition, 2005.

[Bla17] P. Blanchard. *Fast hierarchical algorithms for the low-rank approximation of matrices with applications to materials physics, geostatistics and data analysis*. PhD thesis, University of Bordeaux, 2017.

[BPC+12] H. M. Bik, D. L. Porazinska, S. Creer, J. G. Caporaso, R. Knight, and W. K. Thomas. Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology and Evolution,*, 27:233–243, 2012.

[CC01] T.F. Cox and M. A. A. Cox. *Multidimensional Scaling - Second edition*, volume 88 of *Monographs on Statistics and Applied Probability*. Chapman & al., 2001.

[CKS+10] J. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, and F. D. & al. Bushman. Qiime allows analysis of high-throughput community sequencing data. *Nature Methods*, 7:335–336, 2010.

[DBC+14] K. A. Dafforn, D. J. Baird, A. A. Chariton, M. Y. Sun, M. V. Brown, S. L. Simpson, B. P. Kelaher, and E. L. Johnston. Faster, higher and stronger? the pros and cons of molecular faunal data for assessing ecosystem condition. *Advances in Ecological Research*, 51:1–40, 2014.

[Edg10] R. C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26:2460–2461, 2010.

[Fel04] J. Felsenstein. *Inferring phylogenies*. Sinauer, 2004.

[FRB+16] J. M. Frigerio, F. Rimet, A. Bouchez, E. Chancerel, P. Chaumeil, F. Salin, S. Thérond, M. Kahlert, and A. Franc. diagno-syst: a tool for accurate inventories in metabarcoding. *ArXiv preprint*, arXiv:1611.09410, 2016.

[Gas00] K. J. Gaston. Global patterns in biodiversity. *Nature*, 405:220–227, 2000.

[GN02] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, 2002.

[Gus97] D. Gusfield. *Algorithms on strings, trees, and sequences*. Cambridge University Press, Cambridge, UK, 1997.

[HCBd03] P. D. N. Hebert, A. Cywinska, S. L. Ball, and J. R. deWaard. Biological identifications through dna barcodes. *Proc. R. Soc. Lond. B*, 270:313–321, 2003.

[Hey95] V. H. Heywood. *Global Biodiversity Assessment*. Cambridge University Press, 1995.

[HFSa09] P. M. Hollingsworth, L.L. Forrest, J. L. Spouge, and Hajibabaei & al. A dna barcode for land plants. *PNAS*, 106:12794–12797, 2009.

[HMM96] D. M. Hillis, C. Moritz, and B. Mable. *Molecular Systematics*. Sinauer, Sunderland, Mass., 1996.

[HMT11] N. Halko, P.-G. Martinsson, and J.A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[HSZ+11] M. Hajibabaei, S. Shokralla, X. Zhou, G. A. C. Singer, and D. J. Baird. Environmental barcoding: a next generation sequencing approach for biomnitoring applications using river benthos. *PLoS One*, 6(4):e17497, 2011.

[Ize08] A. J. Izenman. *Modern Multivariate Statistical Techniques*. Springer, NY, 2008.

[JDA+14] S. Joly, T. J. Davies, A. Archambault, A. Bruneau, A. Derry, S. W. Kembel, Peres-Neto P., J. Vamosi, and T. A. Wheeler. Ecology in the age of dna barcoding: the resource, the promise and the challenges ahead. *Molecular Ecology*, 14:221–232, 2014.

[KFR+13] L. Kermarrec, A. Franc, F. Rimet, P. Chaumeil, J.-F. Humbert, and A. Bouchez. Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. *Molecular Ecology Resources*, 13:607–619, 2013.

[KFR+14] L. Kermarrec, A. Franc, F. Rimet, P. Chaumeil, J.-M. Frigerio, J.-F. Humbert, and A. Bouchez. A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshwater Science*, 33:349–363, 2014.

[Kru64] J. B. Kruskal. Multidimensional Scaling by optimizing goodness of fit to a non metric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[Lev92] S. A. Levin. The problem of pattern and scale in ecology: the Robert H. MacArthur Award Lecture. *Ecology*, 73(6):1943–1967, 1992.

[LGRVPAM01] P. López-García, F. Rodriguez-Valera, C. Pedros-Alio, and D. Moreira. Unexpected diversity of small eukaryotes in deep-sea antarctic plankton. *Nature*, 409:603–607, 2001.

[LLMM14] L. Liberti, C. Lavor, N. Maculan, and A. Mucherino. Euclidean distance geometry and applications. *SIAM review*, 56(1):3–69, 2014.

[LV07] J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer, NY, 2007.

[Man99] D. G. Mann. The species concept in diatoms. *Phycologia*, 38(6):437–495, 1999.

[Mar03] A. Margurran. *Measuring Biological Diversity*. Blackwell Publishing, 2003.

[May82] E. Mayr. *The Growth of Biological Thought: Diversity, Evolution and Inheritance.* Harvard University Press, 1982.

[MKB79] K. V. Mardia, J.T. Kent, and J. M. Bibby. *Multivariate Analysis*. Probability and Mathematical Statistics. Academic Press,, 1979.

[MRQ+14] F. Mahé, T Rognes, C. Quince, C. de Vargas, and M. Dunthorn. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, 2:e593:https://doi.org/10.7717/peerj.593, 2014.

[Mur12] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[Mü13] D. Müllner. fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. *Journal of Statistical Software*, 53(9):1–18, http://www.jstatsoft.org/v53/i09/, 2013.

[PAAe12] J. Pawlowski, S. Audic, S. Adl, and *et al.* CBOL Protist Working Group: Barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PloS Biology*, 10(11):e1001419., 2012.

[Pla05] John Platt. Fastmap, metricmap, and landmark mds are all nystrom algorithms. In *AISTATS*, 2005.

[PLE14] J. Pawlowski, F. Lejzerowicz, and P. Eslin. Next-generation environmental diversity surveys of foraminifera: Preparing the future. *Biol. Bull.*, 227:93–106, 2014.

[RCK+16] F. Rimet, P. Chaumeil, F. Keck, L. Kermarrec, V. Vasselon, M. Kahlert, A. Franc, and A. Bouchez. R-syst::diatom: An open-access and curated barcode database for diatoms and freshwater monitoring. *Database: The Journal of Biological Databases and Curation.*, pages 1–21, doi:10.1093/database/baw016, 2016.

[RM00] R. E. Ricklefs and G. L. Miller. *Ecology - 4th edition*. Freeman, 2000.

[SMH+06] M. L. Sogin, H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl. Microbial diversity in the deep sea and the underexplored "rare biosphere". *PNAS*, 103(32):12115–12120, 2006.

[Sor97] Danny C. Sorensen. Implicitly restarted arnoldi/lanczos methods for large scale eigenvalue calculations. *Parallel Numerical Algorithms*, pages 119–165, 1997.

[SS73] R. H. A. Sneath and R. R. Sokal. *Numerical taxonomy*. Freeman, San Francisco, 1973.

[SW81] P. D. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197, 1981.

[SWR+09] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, and E. B. et al. (2009) Hollister. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, 75:7537–7541, 2009.

[Sze11] R. Szeliski. *Computer Vision*. Texts in Computer Science. Sprin, 2011.

[TCHR12] P. Taberlet, E. Coissac, M. Hajibabaei, and L. Rieseberg. Environmental DNA. *Molecular Ecology*, 2:1789–1793., 2012.

[Tor52] W. S. Torgerson. Multidimensional Scaling: I. Theory and Method. *Psychometrika*, 17(4):401–419, 1952.

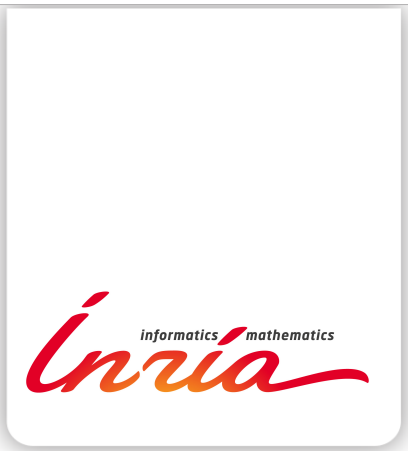
[Vem04] S. S. Vempala. *The Random Projection Method*, volume 65 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Sciences*. American Mathematical Society, 2004.

[vL07] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[Wan12] J. Wang. *Geometric structure if high-dimensional data and dimensionality reduction*. Springer & Higher Education Press, 2012.

[Woo14] David P Woodruff. Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*, 2014.

[Yan06] Z. Yang. *Computational Molecular Evolution*. Oxford Series in Ecology and Evolution. Oxford University Press, 2006.



RESEARCH CENTRE
BORDEAUX – SUD-OUEST

200 avenue de la Vieille Tour
33405 Talence Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399