Bioinformatics, 34(6), 2018, 1009–1015
doi: 10.1093/bioinformatics/btx682
Advance Access Publication Date: 25 October 2017
Original Paper

OXFORD

Data and text mining

Unsupervised multiple kernel learning for heterogeneous data integration

Jérôme Mariette* and Nathalie Villa-Vialaneix*

MIAT, Université de Toulouse, INRA, 31326 Castanet-Tolosan, France

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on May 18, 2017; revised on October 8, 2017; editorial decision on October 22, 2017; accepted on October 24, 2017

Abstract

Motivation: Recent high-throughput sequencing advances have expanded the breadth of available omics datasets and the integrated analysis of multiple datasets obtained on the same samples has allowed to gain important insights in a wide range of applications. However, the integration of various sources of information remains a challenge for systems biology since produced datasets are often of heterogeneous types, with the need of developing generic methods to take their different specificities into account.

Results: We propose a multiple kernel framework that allows to integrate multiple datasets of various types into a single exploratory analysis. Several solutions are provided to learn either a consensus meta-kernel or a meta-kernel that preserves the original topology of the datasets. We applied our framework to analyse two public multi-omics datasets. First, the multiple metagenomic datasets, collected during the *TARA* Oceans expedition, was explored to demonstrate that our method is able to retrieve previous findings in a single kernel PCA as well as to provide a new image of the sample structures when a larger number of datasets are included in the analysis. To perform this analysis, a generic procedure is also proposed to improve the interpretability of the kernel PCA in regards with the original data. Second, the multi-omics breast cancer datasets, provided by The Cancer Genome Atlas, is analysed using a kernel Self-Organizing Maps with both single and multi-omics strategies. The comparison of these two approaches demonstrates the benefit of our integration method to improve the representation of the studied biological system.

Availability and implementation: Proposed methods are available in the R package **mixKernel**, released on CRAN. It is fully compatible with the **mixOmics** package and a tutorial describing the approach can be found on **mixOmics** web site http://mixomics.org/mixkernel/.

Contact: jerome.mariette@inra.fr or nathalie.villa-vialaneix@inra.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recent high-throughput sequencing advances have expanded the breadth of available omics datasets from genomics to transcriptomics, proteomics and methylomics. The integrated analysis of multiple datasets obtained on the same samples has allowed to gain important insights in a wide range of applications from microbial communities profiling (Guidi *et al.*, 2016) to the characterization of molecular signatures of human breast tumours (The Cancer Genome Atlas Network, 2012). However, multiple omics integration analyses remain a challenging task, due to the complexity of biological systems, heterogeneous types (continuous data, counts, factors, networks…) between omics and additional information related to them and the high-dimensionality of the data.

In the literature, several strategies have been proposed to analyse multi-omics datasets. Multivariate approaches are a widely used framework to deal with such problems and several such methods [including partial least squares (PLS) and Canonical Correlation Analysis (CCA)] are provided in the R package **mixOmics** Lê Cao *et al.* (2009).

Similarly, the multiple co-inertia analysis (Chen *et al.*, 2014), use the variability both within and between variables to extract the linear relationships that best explain the correlated structure across datasets. However, these approaches are restricted to the analysis of continuous variables and thus are not generic in term of data types used as inputs. Some works use a case-by-case approach to integrate non numeric information into the analysis: Zhuang *et al.* (2011b) propose a joint non-negative matrix factorization framework to integrate expression profiles to interaction networks by adding network-regularized constraints with the help of graph adjacency matrices and Pavoine *et al.* (2004) and Dray *et al.* (2014) propose extensions of the widely used Principal component analysis (PCoA) approach to integrate information about phylogeny and environmental variables. Finally, some authors propose to use a transformation of all the input datasets into a unified representation before performing an integrated analysis: Kim *et al.* (2012) transforms each data types into graphs, which can be merged before being analysed by standard graph measures and graph algorithms. However, graph based representation are a very constraining and rough way to represent a complex and large dataset.

In this work, we take advantage of the kernel framework to propose a generic approach that can incorporate heterogeneous data types as well as external information in a generic and very flexible way. More precisely, any dataset is viewed through a kernel that provides pairwise information between samples. Kernels are a widely used and flexible method to deal with complex data of various types: they can be obtained from β-diversity measures (Bray and Curtis, 1957; Lozupone *et al.*, 2007) to explore microbiome datasets. They can also account for datasets obtained as read counts by the discrete Poisson kernel (Canale and Dunson, 2011) and are also commonly adopted to quantifies genetic similarities by the state kernel (Kwee *et al.*, 2008; Wu *et al.*, 2010). Our contribution is to propose three alternative approaches able to combine several kernels into one meta-kernel in an unsupervised framework. If multiple kernel approaches are widely developed for supervised analyses, unsupervised approaches are less easy to handle, because no clear *a priori* objective is available. However, they are required to use kernel in exploratory analyses that are the first step to any data analysis.

To evaluate the benefits of the proposed approach, two datasets have been analysed. The first one is the multiple metagenomic dataset collected during the *TARA* Oceans expedition (Bork *et al.*, 2015; Karsenti *et al.*, 2011) and the second one is based on a multi-omic dataset on breast cancer (The Cancer Genome Atlas Network, 2012). A method to improve the interpretability of kernel-based exploratory approaches is also presented and results show that not only our approach allows to retrieve the main conclusions stated in the different papers in a single and fast analysis, but that it can also provide new insights on the data and the typology of the samples by integrating a larger number of information.

2 Materials and methods

2.1 Unsupervised multiple kernel learning

2.1.1 Kernels and notations

For a given set of observations $(x_i)_{i=1,\dots,N}$, taking values in an arbitrary space $\mathcal{X}$, we call 'kernel' a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that provides pairwise similarities between the observations: $K_{ij} := K(x_i, x_j)$. Moreover, this function is assumed to be symmetric ($K_{ij} = K_{ji}$) and positive ($\forall\, n \in \mathbb{N}, \forall\, (\alpha_i)_{i=1,\dots,n} \subset \mathbb{R}, \forall\, (x_i)_{i=1,\dots,n} \subset \mathcal{X}, \sum_{i,i'=1}^{n} \alpha_i \alpha_{i'} K_{ii'} \geq 0$). According to Aronszajn (1950), this ensures that K is the dot product in a uniquely defined Hilbert space $(\mathcal{H}, \langle .,. \rangle)$ of the images of $(x_i)_i$ by a uniquely defined feature map $\phi : \mathcal{X} \to \mathcal{H}$: $K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$. In the sequel, the notation K will be used to denote either the kernel itself or the evaluation matrix $(K_{ij})_{i,j=1,\dots,N}$ depending on the context.

This setting allows us to deal with multiple source datasets in a uniform way, provided that a relevant kernel can be calculated from each dataset (examples are given in Section 3.1 for standard numeric datasets, phylogenetic tree ...). Suppose now that M datasets $(x_i^m)_{i=1,\dots,N}$ (for $m = 1, \dots, M$) are given instead of just one, all obtained on the same samples $i = 1, \dots, N$. M different kernels $(K^m)_{m=1,\dots,M}$ provide different views of the datasets, each related to a specific aspect.

Multiple kernel learning (MKL) refers to the process of linearly combining the M given kernels into a single kernel K^*:

$$K^* = \sum_{m=1}^{M} \beta_m K^m \quad \text{subject to} \begin{cases} \beta_m \geq 0, \forall\, m = 1, \dots, M \\ \sum_{m=1}^{M} \beta_m = 1 \end{cases}. \qquad (1)$$

By definition, the kernel K^* is also symmetric and positive and thus induces a feature space and a feature map (denoted by ϕ^* in the sequel). This kernel can thus be used in subsequent analyses [support vector machine (SVM), kernel PCA (KPCA), kernel self-organizing map (KSOM) ...] as it is supposed to provide an integrated summary of the samples.

A simple choice for the coefficients β_m is to set them all equal to $1/M$. However, this choice treats all the kernels similarly and does not take into account the fact that some of the kernels can be redundant or, on the contrary, atypical. Sounder choices aim at solving an optimization problem so as to better integrate all informations. In a supervised framework, this mainly consists in choosing weights that minimize the prediction error (Gönen and Alpaydin, 2011). For clustering, a similar strategy is used in Zhao *et al.* (2009), optimizing the margin between the different clusters. However, for other unsupervised analyses (such as exploratory analysis, KPCA for instance), such criteria do not exist and other strategies have to be used to choose relevant weights.

As explained in Zhuang *et al.* (2011a), propositions for unsupervised MKL (UMKL) are less numerous than the ones available for the supervised framework. Most solutions (see, e.g. Lin *et al.*, 2010; Zhuang *et al.*, 2011a) seek at providing a kernel that minimizes the distortion between all training data and/or that minimizes the approximation of the original data in the kernel embedding. However, this requires that the datasets $(x_i^m)_i$ ($m = 1, \dots, M$) are standard numerical datasets: the distortion between data and the approximation of the original data are then directly computed in the input space (which is $\mathbb{R}^d$) using the standard Euclidean distance as a reference. Such a method is not applicable when the input dataset is not numerical (i.e. is a phylogenetic tree for instance) or when the different datasets $(x_i^m)_i$ ($m = 1, \dots, M$) do not take value in a common space.

In the sequel, we propose two solutions that overcome this problem: the first one seeks at proposing a consensual kernel, which is the best consensus of all kernels. The second one uses a different point of view and, similarly to what is suggested in Zhuang *et al.* (2011a), computes a kernel that minimizes the distortion between all training data. However, this distortion is obtained directly from the M kernels, and not from an Euclidean input space. Moreover, it is used to provide a kernel representation that preserves the original data topology. Two variants are described: a sparse variant, which also selects the most relevant kernels, and a non sparse variant, when the user does not want to make a selection among the M kernels.

2.1.2 A consensus multiple kernel

Our first proposal, denoted by STATIS-UMKL, relies on ideas similar to STATIS (L'Hermier Des Plantes, 1976; Lavit *et al.*, 1994).

STATIS is an exploratory method designed to integrate multi-block datasets when the blocks are measured on the same samples. STATIS finds a consensus matrix, which is obtained as the matrix that has the highest average similarity with the relative positions of the observations as provided by the different blocks. We propose to use a similar idea to learn a consensus kernel.

More precisely, a measure of similarity between kernels can be obtained by computing their cosines (Cosines are usually preferred over the Frobenius dot product itself because they allow to re-scale the different matrices at a comparable scale. It is equivalent to using the kernel $\tilde{K}^m = \frac{K^m}{||K^m||_F}$ instead of K^m.) according to the Frobenius dot product: $\forall\, m, m' = 1, \ldots, M$,

$$C_{mm'} = \frac{\langle K^m, K^{m'}\rangle_F}{||K^m||_F||K^{m'}||_F} = \frac{\text{Trace}(K^m K^{m'})}{\sqrt{\text{Trace}\left((K^m)^2\right)\ \text{Trace}\left((K^{m'})^2\right)}}. \quad (2)$$

$C_{mm'}$ can be viewed as an extension of the RV-coefficient (Robert and Escoufier, 1976) to the kernel framework, where the RV-coefficient is computed between $\left(\phi^m(x_i^m)\right)_i$ and $\left(\phi^{m'}(x_i^{m'})\right)_i$ (where ϕ^m is the feature map associated to K^m).

The similarity matrix $\mathbf{C} = (C_{mm'})_{m,m'=1,\ldots,M}$ provides information about the resemblance between the different kernels and can be used as such to understand how they complement each other or if some of them provide an atypical information. It also gives a way to obtain a summary of the different kernels by choosing a kernel K^* which maximizes the average similarity with all the other kernels:

$$\begin{aligned} \text{maximize}_\beta \quad & \sum_{m=1}^M \left\langle K^*_{\mathbf{v}}, \frac{K^m}{||K^m||_F}\right\rangle_F = \mathbf{v}^\top \mathbf{C}\mathbf{v} \\ \text{for } & K^*_{\mathbf{v}} = \sum_{m=1}^M \nu_m K^m \\ \text{and } & \mathbf{v} \in \mathbb{R}^M \text{ such that } ||\mathbf{v}||_2 = 1. \end{aligned} \quad (3)$$

The solution of the optimization problem of Equation (3) is given by the eigen-decomposition of $\mathbf{C}$. More precisely, if $\mathbf{v} = (\nu_m)_{m=1,\ldots,M}$ is the first eigenvector (with norm 1) of this decomposition, then its entries are all positive (because the matrices K^m are positive) and are the solution of the maximization of $\mathbf{v}^\top\mathbf{C}\mathbf{v}$. Setting $\beta = \frac{\mathbf{v}}{\sum_{m=1}^M \nu_m}$ thus provides a solution satisfying the constraints of Equation (1) and corresponding to a consensual summary of the M kernels.

Note that this method is equivalent to performing multiple CCA between the multiple feature spaces, as suggested in Wang *et al.* (2008) in a supervised framework, or in Ren *et al.* (2013) for multiple kernel PCA. However, only the first axis of the CCA is kept and a L^2-norm constrain is used to allow the solution to be obtained by a simple eigen-decomposition. This solution is better adapted to the case where the number of kernels is small.

2.1.3 A sparse kernel preserving the original topology of the data

Because it focuses on consensual information, the previous proposal tends to give more weights to kernels that are redundant in the ensemble of kernels and to discard the information given by kernels that provide complementary informations. However, it can also be desirable to obtain a solution which weights the different images of the dataset provided by the different kernels more evenly. A second solution is thus proposed, which seeks at preserving the original topology of the data. This method is denoted by sparse-UMKL in the sequel.

More precisely, weights are optimized such that the local geometry of the data in the feature space is the most similar to that of the original data. Since the input datasets are not Euclidean and do not take values in a common input space, the local geometry of the original data cannot be measured directly as in Zhuang *et al.* (2011a). It is thus approximated using only the information given by the M kernels. To do so, a graph, the k-nearest neighbour graph (for a given $k \in \mathbb{N}^*$), $\mathcal{G}^m$, associated with each kernel K^m is built. Then, a $(N \times N)$-matrix $\mathbf{W}$, representing the original topology of the dataset is defined such that W_{ij} is the number of times the pair (i, j) is in the edge list of $\mathcal{G}^m$ over $m = 1, \ldots, m$ (i.e. the number of times, over $m = 1, \ldots, M$, that x_i^m is one of the k nearest neighbours of x_j^m or x_j^m is one of the k nearest neighbours of x_i^m).

The solution is thus obtained for weights that ensure that $\phi^*(x_i)$ and $\phi^*(x_j)$ are 'similar' (in the feature space) when W_{ij} is large. To do so, similarly as Lin *et al.* (2010), we propose to focus on some particular features of $\phi^*(x_i)$ which are relevant to our problem and correspond to their similarity (in the feature space) with all the other $\phi^*(x_j)$. More precisely for a given $\beta \in \mathbb{R}^M$, we introduce the N-dimensional vector $\Delta_i(\beta) = \left\langle \phi^*_\beta(x_i), \begin{pmatrix} \phi^*_\beta(x_1) \\ \vdots \\ \phi^*_\beta(x_N) \end{pmatrix} \right\rangle = \begin{pmatrix} K^*_\beta(x_i, x_1) \\ \vdots \\ K^*_\beta(x_i, x_N) \end{pmatrix}$. But, contrary to Lin *et al.* (2010), we do not rely on a distance in the original space to measure topology preservation but we directly use the information provided by the different kernels through $\mathbf{W}$. The following optimization problem is thus solved:

$$\begin{aligned} \text{minimize}_\beta \quad & \sum_{i,j=1}^N W_{ij}\|\Delta_i(\beta) - \Delta_j(\beta)\|^2 \\ & \text{for } K^*_\beta = \sum_{m=1}^M \beta_m K^m \\ & \text{and } \beta \in \mathbb{R}^M \text{ such that } \beta_m \geq 0 \text{ and } \sum_{m=1}^M \beta_m = 1. \end{aligned} \quad (4)$$

The optimization problem of Equation (4) expands as

$$\begin{aligned} \text{minimize}_\beta \quad & \sum_{m,m'=1}^M \beta_m \beta_{m'} S_{mm'} \\ & \text{for } \beta \in \mathbb{R}^M \text{ such that } \beta_m \geq 0 \text{ and } \sum_{m=1}^M \beta_m = 1, \end{aligned} \quad (5)$$

for $S_{mm'} = \sum_{i,j=1}^N W_{ij}\langle \Delta_i^m - \Delta_j^m, \Delta_i^{m'} - \Delta_j^{m'}\rangle$ and $\Delta_i^m = \begin{pmatrix} K^m(x_i, x_1) \\ \vdots \\ K^m(x_i, x_N) \end{pmatrix}$. The matrix $\mathbf{S} = (S_{mm'})_{m,m'=1,\ldots,M}$ is positive and the problem is thus a standard Quadratic Programming (QP) problem with linear constraints, which can be solved by using the R package **quadprog**. Since the constrain $\sum_{m=1}^M \beta_m = 1$ is an L^1 constrain in a QP problem, the produced solution will be sparse: a kernel selection is performed because only some of the obtained $(\beta_m)_m$ are non zero. Although desirable when the number of kernels is large, this property can be a drawback when the number of kernels is small and that using all kernels in the integrated exploratory analysis is expected. To address this issue, a modification of Equation (5) is proposed in the next section.

2.1.4 A full kernel preserving the original topology of the data

To get rid of the sparse property of the solution of Equation (5), an L^2 constrain can be used to replace the L^1 constrain, similarly to Equation (3):

$$\begin{aligned} \text{minimize}_{\mathbf{v}} \quad & \sum_{m,m'=1}^M \nu_m \nu_{m'} S_{mm'} \\ & \mathbf{v} \in \mathbb{R}^M \text{ such that } \nu_m \geq 0 \text{ and } ||\mathbf{v}||_2 = 1, \end{aligned} \quad (6)$$

and to finally set $\beta = \frac{\mathbf{v}}{\sum_m \nu_m}$. This problem is a Quadratically constrained quadratic programme, which is known to be hard to solve.

For a similar problem, Lin *et al.* (2010) propose to relax the problem into a semidefinite programming optimization problem. However, a simpler solution is provided by using alterning direction method of multipliers; Boyd *et al.* (2011). More precisely, the optimization problem of Equation (6) is re-written as

$$\begin{aligned} \text{minimize}_{\mathbf{x}} \text{ and } _{\mathbf{z}} \quad & \mathbf{x}^T\mathbf{S}\mathbf{x} + \mathbb{I}_{\{\mathbf{x}\geq 0\}}(\mathbf{x}) + \mathbb{I}_{\{\mathbf{z}\geq 1\}} \\ \text{such that } & \mathbf{x} - \mathbf{z} = 0 \end{aligned}$$

and is solved with the method of multipliers. Final weights are then obtained by re-scaling the solution $\beta := \frac{\mathbf{z}}{\sum_m z_m}$. The method is denoted by full-UMKL in the sequel.

2.2 KPCA and enhanced interpretability

The combined kernel can be used in subsequent exploratory analyses to provide an overview of the relations between samples through the different integrated datasets. Any method based only on dot product and norm computations can have a kernel version and this includes a variety of standard methods, such as PCA (KPCA, see below), clustering [kernel k-means, Schölkopf *et al.* (2004)] or more sophisticated approaches that combine clustering and visualization like KSOM (Mac Donald and Fyfe, 2000). In this section, we focus on the description of KPCA because it is close to the standard approaches that are frequently used in metagenomics (PCoA) and is thus a good baseline analysis for investigating the advantages of our proposals. Moreover, we have used KPCA to propose an approach that is useful to improve the interpretability of the results. Section 4.2 illustrates that our method is not restricted to this specific analysis and is straightforwardly extensible to other exploratory tools.

2.2.1 Short description of KPCA

KPCA, introduced in Schölkopf *et al.* (1998), is a PCA analysis performed in the feature space induced by the kernel K^*. It is equivalent to standard MDS [i.e. metric MDS or PCoA; Togerson (1958)] for Euclidean dissimilarities. Without loss of generality, the kernel K^* is supposed centered (if K^* is not centered, it can be made so by computing $K^* - \frac{1}{N}K^*\mathbf{I}_N + \frac{1}{N^2}\mathbf{I}_N^\top K^*\mathbf{I}_N$, with $\mathbf{I}_N$ a vector with N entries equal to 1.). KPCA simply consists in an eigen-decomposition of K^*: if $(\boldsymbol{\alpha}_k)_{k=1,\ldots,N} \in \mathbb{R}^N$ and $(\lambda_k)_{k=1,\ldots,N}$ respectively denote the eigenvectors and corresponding eigenvalues (ranked in decreasing order) then the PC axes are, for $k = 1, \ldots, N$, $a_k = \sum_{i=1}^N \alpha_{ki}\phi^*(x_i)$, where $\boldsymbol{\alpha}_k = (\alpha_{ki})_{i=1,\ldots,N}$. $a_k = (a_{ki})_{i=1,\ldots,N}$ are orthonormal in the feature space induced by the kernel: $\forall k, k', \langle a_k, a_{k'}\rangle = \boldsymbol{\alpha}_k^\top K^* \boldsymbol{\alpha}_{k'} = \delta_{kk'}$ with $\delta_{kk'} = \begin{cases} 0 & \text{if } k \neq k' \\ 1 & \text{otherwise} \end{cases}$. Finally, the coordinates of the projections of the images of the original data, $(\phi^*(x_i))_i$, onto the PC axes are given by: $\langle a_k, \phi^*(x_i)\rangle = \sum_{j=1}^N \alpha_{kj}K^*_{ji} = K^*_{i.}\boldsymbol{\alpha}_k = \lambda_k\alpha_{ki}$, where $K^*_{i.}$ is the ith row of the kernel K^*.

These coordinates are useful to represent the samples in a small dimensional space and to better understand their relations. However, contrary to standard PCA, KPCA does not come with a variable representation, since the samples are described by their relations (via the kernel) and not by standard numeric descriptors. PC axes are defined by their similarity to all samples and are thus hard to interpret.

2.2.2 Interpretation

There are few attempts, in the literature, to help understand the relations of KPCA with the original measures. When the input datasets take values in $\mathbb{R}^d$, Reverter *et al.* (2014) propose to add a representation of the variables to the plot, visualizing their influence over the results from derivative computations. However, this approach would make little sense for datasets like ours, i.e. described by discrete counts.

We propose a generic approach that assesses the influence of variables and is based on random permutations. More precisely, for a given measure j, that is used to compute the kernel K^m, the values observed on this measure are randomly permuted between all samples and the kernel is re-computed: $\tilde{K}^{m,j}$. For species abundance datasets, the permutation can be performed at different phylogeny levels, depending on the user interest. Then, using the weights found with the original (non permuted) kernels, a new meta-kernel is obtained $\tilde{K}^* = \sum_{l\neq m}\beta_l K^l + \beta_m \tilde{K}^{m,j}$. The influence of the measure j on a given PC subspace is then assessed by computing the Crone-Crosby distance (Crone and Crosby, 1995) at the axis level: $\forall k = 1, \ldots, N$, $D_{\text{cc}}(\alpha_k, \tilde{\alpha}_k) = \frac{1}{\sqrt{2}}||\alpha_k - \tilde{\alpha}_k||$, where α_k and $\tilde{\alpha}_k$ respectively denote the eigenvectors of the eigen-decomposition of K^* and $\tilde{K}^*$ (Note that a similar distance can be computed at the entire projection space level but, since axes are naturally ordered in PCA, we chose to restrict to axis-specific importance measures).

Finally, the KPCA interpretation is done similarly as for a standard PCA: the interpretation of the axes $(a_k)_{k=1,\ldots,N}$ is done with respect to the observations $(x_i)_{i=1,\ldots,N}$ which contribute the most to their definition, when important variables are the ones leading to the largest Crone-Crosby distances.

Methods presented in the paper are available in the R package **mixKernel**, released on CRAN. Further details about implemented functions are provided in Supplementary Section S1.

3 Case studies

3.1 *TARA* oceans

The *TARA* Oceans expedition (Bork *et al.*, 2015; Karsenti *et al.*, 2011) facilitated the study of plankton communities by providing oceans metagenomic data combined with environmental measures to the scientific community. During the expedition, 579 samples were collected for morphological, genetic and environmental analyses, from 75 stations in epipelagic and mesopelagic waters across eight oceanic provinces. The *TARA* Oceans consortium partners analysed prokaryotic (Sunagawa *et al.*, 2015), viral (Brum *et al.*, 2015) and eukaryotic-enriched (de Vargas *et al.*, 2015) size fractions and provided an open access to the raw datasets and processed materials. So far, all articles related to *TARA* Oceans that aim at integrating prokaryotic, eukaryotic and viral communities, took advantage of the datasets only by using co-occurrence associations Guidi *et al.* (2016), Lima-Mendez *et al.* (2015) and Villar *et al.* (2015). The integration analysis of the whole material aims at providing a more complete overview of the relations between all collected informations.

A total of 48 selected samples were collected in height different oceans or seas: Indian Ocean, Mediterranean Sea, North Atlantic Ocean, North Pacific Ocean, Red Sea, South Atlantic Ocean, South Pacific Ocean (SPO) and South Ocean (SO). Using these samples, eight (dis)similarities were computed using public preprocessed datasets, which are all available from the *TARA* Oceans consortium partner's websites. These dissimilarities provide information about environmental variables, phylogenetic similarities, prokaryotic functional processes, different aspects of the eukaryotic dissimilarities and virus composition. Selected datasets as well as chosen kernels are fully described in Supplementary Section S2. The meta-kernel was analysed using a KPCA and the most important variables were assessed as described in Section 2.2.2.

3.2 The Cancer Genome Atlas

The Cancer Genome Atlas Network (2012) (TCGA) provides multi-omics datasets from different tumour types, such as colon, lung and breast cancer. In this work, we consider normalized and pre-filtered breast cancer datasets available from the **mixOmics** website (http://mixomics.org/tcga-example/). Using the 989 available samples, three kernels were computed. The **TCGA.mRNA** kernel provides a gene expression dissimilarity measure computed on the expression of 2000 mRNAs, the **TCGA.miRNA** describes the expression of 184 miRNAs and the methylation aspect is assessed by the **TCGA.CpG** kernel, computed on 2000 CpG probes. These three kernels were obtained using the Gaussian kernel, $K_{ij} = e^{-\sigma||x_i - x_j||^2}$ with σ equal to the median of $\left\{\frac{1}{||x_i - x_j||^2}\right\}_{i<j}$. The combined kernel was used in KSOM that has been implemented with the R package **SOMbrero** (Boelaert *et al.*, 2014). KSOM is an exploratory tool that combines clustering and visualization by mapping all samples onto a 2D (generally squared) grid made of a finite number of units (also called clusters). It has been shown relevant, e.g. to provide a relevant taxonomy of Amazonian butterflies from DNA barcoding in (Olteanu and Villa-Vialaneix, 2015).

The results of our analysis with the combined kernel were compared with the results obtained with a simple analysis that uses only one of the kernel. The comparison was performed using a quality measure specific to SOM, the *topographic error* (TE), which is the ratio of the second best matching unit that falls in the direct neighbour, on the grid, of the chosen unit over all samples (Pölzlbauer, 2004). In addition, breast cancer subtypes, i.e. *Basal*, *Her2*, *LumA* or *LumB* are provided for every sample and were used as an *a priori* class to compute clustering quality measures (they were thus excluded from the exploratory analysis). More precisely, (i) the average cluster purity, i.e. the mean over all clusters on the grid of the frequency of the majority vote cancer subtype and (ii) the normalized mutual information (NMI) (Danon *et al.*, 2005) between cancer subtype and cluster, which is a value comprised between 0 and 1 (1 indicating a perfect matching between the two classifications).

4 Results and discussion

Sections 4.1 and 4.2 provide and discuss results of exploratory analyses performed from the two sets of datasets described in the previous section. More precisely, Section 4.1 explores the datasets studied in Sunagawa *et al.*, (2015), Brum *et al.* (2015) and de Vargas *et al.* (2015) with KPCA. This illustrates how a multiple metagenomic dataset can be combined with external information to provide an overview of the structure of the different samples. In addition, Section 4.2 shows that our approach is not restricted nor to metagenomic neither to KPCA by studying the multi-omic dataset related to breast cancer with KSOM.

All analyses presented in this section use the full-UMKL strategy. However, for both datasets, a study of the correlation between kernels in the line of the STATIS-UMKL approach is provided in Supplementary Section S4 and shows how this approach helps understand the relations between the multiple datasets. Moreover, a comparison between the different multiple kernel strategies is discussed in Supplementary Section S5 and justifies the choice of full-UMKL for our problems. All combined kernels have been implemented with our package **mixKernel**, as well as all KPCA results.

4.1 Exploring *TARA* oceans datasets with a single KPCA

In a preliminary study (fully reported in Supplementary Section S3), an exploratory analysis was performed using a KPCA with only the three *TARA* Oceans datasets studied in Sunagawa *et al.* (2015) and the full-UMKL strategy. The results show that the main variability between the different samples is explained similarly as in Sunagawa *et al.* (2015): the most important variables returned by our method are those discussed in this article to state the main conclusions.

A further step is then taken by integrating all *TARA* Oceans datasets described in Supplementary Section S2. Supplementary Section S4.1 shows that **pro.phylo** and **euk.pina** are the most correlated kernels to environmental and physical variables, unlike large organism size fractions, i.e. **euk.meso** which is strongly geographically structured. Figure 1 (left) displays the projection of the samples on the first two axes of the KPCA. Figure 1 (right) and Supplementary Figure S16 provide the five most important variables for each datasets, respectively for the first and the second axes of the KPCA. To obtain these figures, abundance values were permuted at 56 prokaryotic phylum levels for the **pro.phylo** kernel, at 13 eukaryotic phylum levels for **euk.pina**, **euk.nano**, **euk.micro** and **euk.meso** and at 36 virus family levels for the **vir.VCs** kernel. Variables used for **phychem** and **pro.NOGs** were the same than the one used in the restricted analysis. Additionally, the explained variance supported by the first 15 axes is provided in Supplementary Figure S17. Using an R implementation of the methods on a one CPU computer with 16 GB memory, the computational cost to combine the three kernels is only ∼3 s. Permutations to assess the eight kernels important variables are computationally much more demanding if performed at a fine level as we did. In our case, they took ∼3 min.

Contrary to the restricted analysis, Figure 1 does not highlight any particular pattern in terms of depth layers but it does in terms of geography. SO samples are gathered in the bottom-center of the KPCA projection and SPO samples are gathered on the top-left side. Figure 1 shows that the most important variables come from the **phychem** kernel (especially the longitude) and from kernels representing the eukaryotic plankton. More specifically, large size organisms are the most important: *rhizaria* phylum for **euk.meso** and *alveolata* phylum for **euk.nano**. The abundance of *rhizaria* organisms also ranks first between important variables of the second KPCA axis, followed by the *opisthokonta* phylum for **euk.nano**. The display of these variables on the KPCA projection reveals a gradient on the first axis for both the *alveolata* phylum abundance (Supplementary Fig. S18) and the longitude (Supplementary Fig. S19) and on the second axis for *rhizaria* (Supplementary Fig. S20) and *opisthokonta* (Supplementary Fig. S21) abundances. This indicates that SO and SPO epipelagic waters mainly differ in terms of *Rhizarians* abundances and both of them differ from the other studied waters in terms of *alveolata* abundances.

The integration of *TARA* Oceans datasets shows that the variability between epipelagic samples is mostly driven by geography rather than environmental factors and that this result is mainly explained by the strong geographical structure of large eukaryotic communities. Studied samples were all collected from epipelagic layers, where water temperature does not vary much, which explains the poor influence of the prokaryotic dataset in this analysis.

4.2 Clustering breast cancer multi-omics datasets

KSOM was used to obtain a map from the three datasets presented in Section 3.2: mRNAs, miRNAs and methylation datasets. The results were compared with a single-omic analysis with the same method (KSOM). KSOM maps were trained with a 5×5 grid, 5000 iterations in a stochastic learning framework and a Gaussian neighbourhood controlled with the Euclidean distance between units on

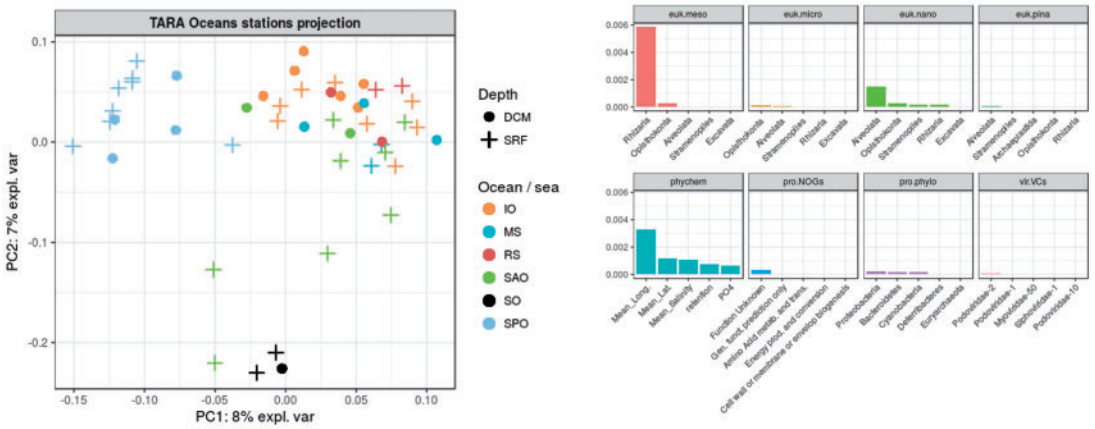

Fig. 1. Left: Projection of the observations on the first two KPCA axes. Colors represent the oceanic regions and shapes the depth layers. Right: The five most important variables for each of the eight datasets, ranked by decreasing Crone-Crosby distance

Table 1. Performance results (average cluster purity and NMI with respect to the cancer subtype) of KSOM (average over 100 maps and SD between parenthesis) for the three single-omics datasets: **TCGA.mRNA**, **TCGA.miRNA** and **TCGA.CpG** and the meta-kernel, **TCGA.all**

	TCGA.mRNA	TCGA.miRNA	TCGA.CpG	TCGA.all
Purity	0.67 (0.02)	0.62 (0.02)	0.66 (0.01)	**0.70** (0.02)
NMI	0.31 (0.02)	0.17 (0.01)	0.18 (0.01)	**0.34** (0.01)

The bold text refers to the best method with the best score.

the grid. The computational cost to combine the kernels was ∼2 min and ∼12 s were required to generate one map.

Table 1 reports KSOM performances obtained over 100 maps (mean and SD) for the three kernels: **TCGA.mRNA**, **TCGA.miRNA** and **TCGA.CpG** and the meta-kernel, denoted **TCGA.all**. Results are reported in terms of average cluster purity and NMI, with respect to the cancer subtypes. All TE were found to be equal to 0. This indicates a good organization of the results on the grid, with respect to the topology of the original dataset as represented in the input kernel. Finally the map with the best NMI obtained for the meta-kernel is given in Figure 2.

For all quality criteria, the integrated analysis gives better results (with respect to cancer subtype) than single-omics analyses (all differences are significant according to a student test, risk 1%). This can be explained by the fact that the information provided especially by mRNA and CpG are complementary, as described in the analysis of correlations between kernels in Supplementary Section S4.2. In addition, Figure 2 shows that the clustering produced by the KSOM is relevant to discriminate between the different breast cancer subtypes and to identify their relations (e.g. subtypes *LumA* and *Basal* are closer to subtypes *LumB* and *Her2* than they are from each other). The organization of cancer subtypes on the map is in accordance with what is reported in (Sørliea *et al.*, 2001) (from cDNA microarray). However, it has been obtained with additional datasets and thus provides a finer typology of samples. It shows that some *LumA* samples are mixed with *LumB* samples (cluster at the bottom left of the map) and that samples classified in the middle of the map probably have an ambiguous type. It also gives clue to define which samples are typical from a given cancer subtype. In addition, Supplementary Section S6 shows the results obtained by KPCA that are consistent with those of KSOM. It also provides a list of features

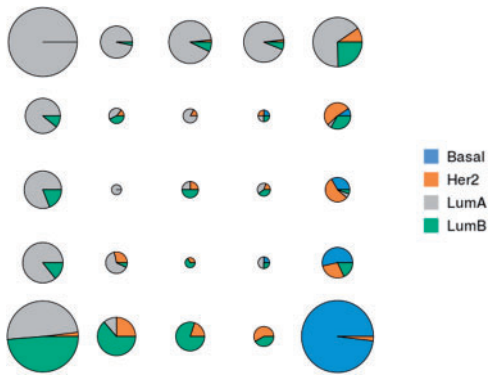

Fig. 2. For each unit of the map, distribution of breast cancer subtypes. Colors represent the different breast cancer subtypes (*Basal*, *Her2*, *LumA* or *LumB*) and the area of the pie charts is proportional to the number of samples classified in the corresponding unit

(mRNA, miRNA and CpG probes) that are potentially interesting to discriminate between breast cancer subtypes.

5 Conclusion

The contributions of this article to the analysis of multi-omics datasets are 2-folds: first, we have proposed three unsupervised kernel learning approaches to integrate multiple datasets from different types, which either allow to learn a consensual meta-kernel or a meta-kernel preserving the original topology of the data. Second, we have improved the interpretability of the KPCA by assessing the influence of input variables in a generic way.

The experiments performed on *TARA* Oceans and breast cancer datasets showed that presented methods allow to give a fast and accurate insight over the different datasets within a single analysis and is able to bring new insights as compared with separated single-dataset analyses. Future work includes the addition of more kernels and post-processing methods for the analysis into our package **mixKernel**.

Acknowledgments

The authors are grateful to Rémi Flamary (University of Nice, France) for helpful suggestions on optimization and to the **mixOmics** team, especially to Kim-Anh Le Cao and Florian Rohart, for helping to integrate our method in the **mixOmics** package. We also thank the three anonymous reviewers for

valuable comments and suggestions which helped to improve the quality of the article.

Conflict of Interest: none declared.

References

Aronszajn,N. (1950) Theory of reproducing kernels. *Trans. Am. Math. Soc.*, **68**, 337–404.

Boelaert,J. *et al.* (2014) SOMbrero: an r package for numeric and non-numeric self-organizing maps. In: Villmann,T., Schleif,F., Kaden,M. and Lange,M. (eds.) *Advances in Self-Organizing Maps and Learning Vector Quantization (Proceedings of WSOM 2014), Volume 295 of Advances in Intelligent Systems and Computing, Pages 219–228, Mittweida, Germany*. Springer Verlag, Berlin, Heidelberg.

Bork,P. (2015) Tara oceans studies plankton at planetary scale. *Science*, **348**, 873–873.

Boyd,S. *et al.* (2010) Distributed optimization and statistical learning via the alterning direction method of multipliers. *Found. Trends Mach. Learn.*, **3**, 1–122.

Bray,R. and Curtis,J. (1957) An ordination of the upland forest communities of southern wisconsin. *Ecol. Monogr.*, **27**, 325–349.

Brum,J. *et al.* (2015) Patterns and ecological drivers of ocean viral communities. *Science*, **348**, 1261498.

Canale,A. and Dunson,D.B. (2011) Bayesian kernel mixtures for counts. *J. Am. Stat. Assoc.*, **106**, 1528–1539.

Chen,M. *et al.* (2014) A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, **15**, 162.

Crone,L. and Crosby,D. (1995) Statistical applications of a metric on subspaces to satellite meteorology. *Technometrics*, **37**, 324–328.

Danon,L. *et al.* (2005) Comparing community structure identification. *J. Stat. Mech.*, **2005**, P09008.

de Vargas,C. *et al.* (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science*, **348**, 1261605.

Dray,S. *et al.* (2015) Considering external information to improve the phylogenetic comparison of microbial communities: a new approach based on constrained double principal coordinates analysis (cDPCoA). *Mol. Ecol. Resources*, **15**, 242–249.

Gönen,M. and Alpaydin,E. (2011) Multiple kernel learning algorithms. *J. Mach. Learn. Res.*, **12**, 2211–2268.

Guidi,L. *et al.* (2016) Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, **532**, 465–470.

Karsenti,E. *et al.* (2011) A holistic approach to marine eco-systems biology. *PLoS Biol.*, **9**, e1001177.

Kim,D. *et al.* (2012) Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J. Biomed.*, **45**, 1191–1198.

Kwee,L. *et al.* (2008) A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Genet.*, **82**,), 386–397.

Lavit,C. *et al.* (1994) The act (statis method). *Comput. Stat. Data Anal.*, **18**, 97–119.

Lê Cao,K. *et al.* (2009) *****Omics: an R package to unravel relationships between two omics data sets. *Bioinformatics*, **25**, 2855–2856.

L'Hermier Des Plantes,H. (1976) *Structuration des tableaux à trios indices de la statistique*. PhD Thesis, Université de Montpellier. Thèse de troisième cycle.

Lima-Mendez,G. *et al.* (2015) Determinants of community structure in the global plankton interactome. *Science*, **348**, 1262073.

Lin,Y. *et al.* (2010) Multiple kernel learning for dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.*, **33**, 1147–1160.

Lozupone,C. *et al.* (2007) Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, **73**, 1576–1585.

Mac Donald,D. and Fyfe,C. (2000) The kernel self organising map. In: *Proceedings of 4th International Conference on knowledge-based Intelligence Engineering Systems and Applied Technologies*, pp. 317–320.

Olteanu,M. and Villa-Vialaneix,N. (2015) On-line relational and multiple relational SOM. *Neurocomputing*, **147**, 15–30.

Pavoine,S. *et al.* (2004) From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *J. Theor. Biol.*, **228**, 523–537.

Pölzlbauer,G. (2004) Survey and comparison of quality measures for self-organizing maps. In: Paralic, J., Pölzlbauer, G. and Rauber, A. (eds.) *Proceedings of the Fifth Workshop on Data Analysis (WDA'04)*, pp. 67–82, Sliezsky dom, Vysoke Tatry, Slovakia. Elfa Academic Press.

Ren,S. *et al.* (2013) Multi-kernel PCAwith discriminant manifold for hoist monitoring. *J. Appl. Sci.*, **13**, 4195–4200.

Reverter,F. *et al.* (2014) Kernel-PCA data integration with enhanced interpretability. *BMC Syst. Biol.*, **8**, S6.

Robert,P. and Escoufier,Y. (1976) A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Appl. Stat.*, **25**, 257–265.

Schölkopf,B. *et al.* (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, **10**, 1299–1319.

Schölkopf,B. *et al.* (2004) *Kernel Methods in Computational Biology*. MIT Press, London, UK.

Sørlie,T. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA*, **98**, 10869–10874.

Sunagawa,S. *et al.* (2015) Structure and function of the global ocean microbiome. *Science*, **348**, 1261359.

The Cancer Genome Atlas Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **498**, 61–70.

Togerson,W. (1958) *Theory and Methods of Scaling*. Wiley, New York, NY.

Villar,E. *et al.* (2015) Environmental characteristics of agulhas rings affect interocean plankton transport. *Science*, **348**, 1261447.

Wang,Z. *et al.* (2008) MultiK-MHKS: a novel multiple kernel learning algorithm. *IEEE Trans. n Pattern Anal. Mach. Intell.*, **30**, 348–353.

Wu,M. *et al.* (2010) Powerful snp-set analysis for case-control genomewide association studies. *Am. J. Hum. Genet.*, **86**, 929–942.

Zhao,B. *et al.* (2009) Multiple kernel clustering. In: Apte, C., Park, H., Wang,K. and Zaki, M. (eds.) *Proceedings of the 2009 SIAM International Conference on Data Mining (SDM), pages 638–649*. SIAM, Philadelphia, PA.

Zhuang,J. *et al.* (2011a) Unsupervised multiple kernel clustering. In: *Journal of Machine Learning Research: Workshop and Conference Proceedings*, Vol. 20, pp. 129–144, Taoyuan, Taiwain.

Zhuang,S. *et al.* (2011b) Anovel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules. *Bioinformatics*, **27**, i401–i409.