Kernel Approaches for Multi-Omics Data Analysis and Biomarker Discovery

MSCA-ITN-2020 European Training Network EMUSE

Mitja Briscik

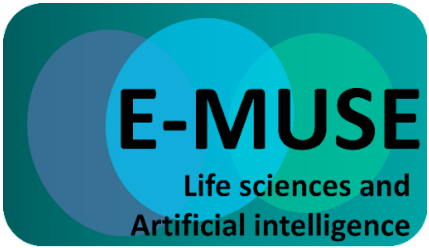
December 12, 2024

Contents

Complex microbial Ecosystems MUltiScale modElling: mechanistic and data driven approaches integration.

Combining **artificial intelligence** and **systems biology** :

- develop innovative modelling methodologies
- improve knowledge about complex biological systems
- predict their evolution



Application of the results to macro-scale properties related to **cheese ripening** and **consumer preference**.

PhD fellowship in development kernel approaches for the integration of **biological data** from heterogeneous sources

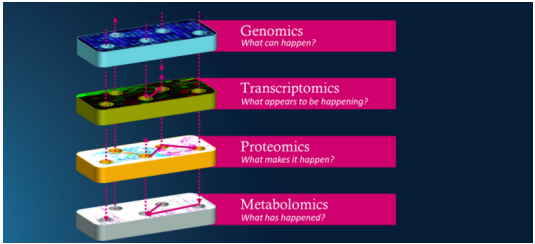
Biological data

- Multi-omics datasets have become more and more available

PhD fellowship in development kernel approaches for the integration of biological data from **heterogeneous sources**

Heterogeneous sources

- Systems biology often produces datasets of heterogeneous types (continuous data, counts, factors, networks . . .) types



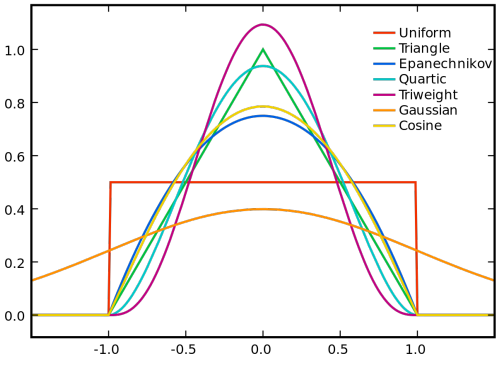
PhD fellowship in development **kernel approaches** for the integration of biological data from heterogeneous sources

What is a kernel?

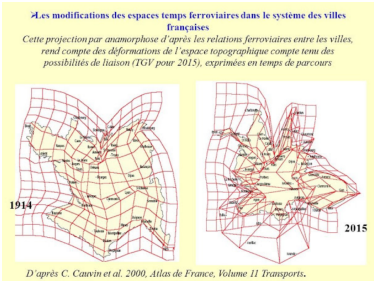
A function k defined as k: $\chi \times \chi \longrightarrow \mathbb{R}$ s.t.

- $k(x_i, x_j) = k(x_j, x_i)$
- c'$\boldsymbol{K}$c $\geqslant 0$ $\forall c \in \mathbb{R}$

where $\boldsymbol{K}$ is the $n \times n$ matrix containing all the data pairwise similarities $\boldsymbol{K} = k(x_i, x_j)$.



Linearity is the biggest advantage of most matrix factorization methods, but it comes at the cost of a substantial loss of explanatory power. Nonlinear alternatives, such as deep generative models in the form of variational autoencoders, have proven to be powerful generalizations of factor analysis and have been successfully applied to a variety of single-cell genomics technologies, albeit at the cost of reduced interpretability (Argelaguet et al., 2021).

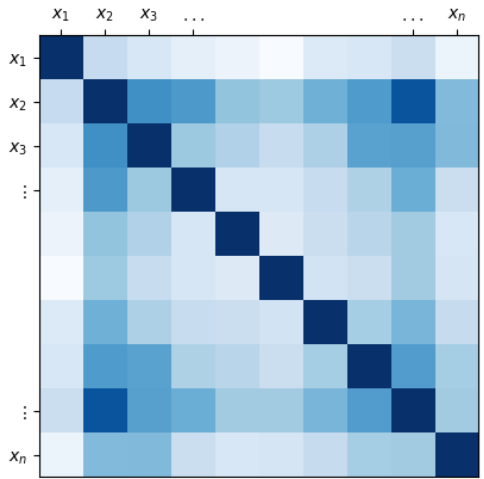


PhD fellowship in development **kernel approaches** for the **integration** of biological data from heterogeneous sources

Why?

Any dataset is viewed through a kernel fuction, that provides pairwise information between samples contained in $\boldsymbol{K}$

- Analyze multiple heterogeneus sources datasets in uniform way
- Account for nonlinearity in the data



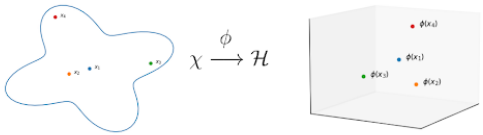
A positive definite kernel is identical to a dot product in another space, the **feature space**

Kernel trick

Any dataset is viewed through a kernel function, that provides pairwise information between samples contained in $\boldsymbol{K}$

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

It allows to perform operations implicitly in the feature space.

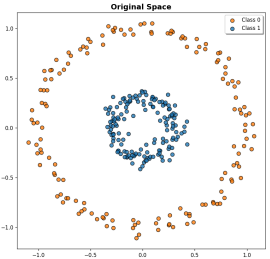


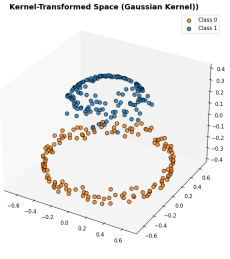
Transform linear methods into **nonlinear methods**

Kernelization

- Replacing the dot product by a general kernel

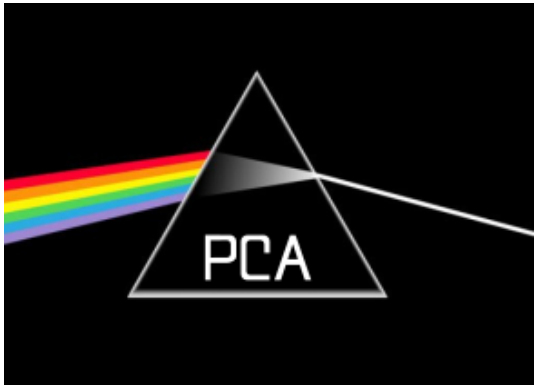
The algorithm remains identical as the computational cost.





Contents

- **Dimensionality Reduction**: PCA transforms data into a lower-dimensional space while maximazing the explained variance.
- **Orthogonal Components**: Identifies uncorrelated principal components ranked by the amount of explained variance.



PCA

Given a set of centered observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ with $\boldsymbol{x}_i \in \mathbb{R}^p$, PCA diagonalizes the covariance matrix

$$C = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \tag{1}$$

with the eigenvalue equation

$$\lambda \boldsymbol{v} = C\boldsymbol{v} \tag{2}$$

with $\lambda \geq 0$ the eigenvalues of C with $\boldsymbol{v}$ the corresponding eigenvectors, $\boldsymbol{v} \in \mathbb{R}^p$

KPCA

Given a set of centered observations in the feature space i.e. $\sum_{j=1}^{n} \phi(\boldsymbol{x}_i) = 0$, the covariance matrix is diagonalized in the feature space

$$\tilde{C} = \frac{1}{n} \sum_{i=1}^{n} \phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i)^T \tag{3}$$

with the eigenvalue equation

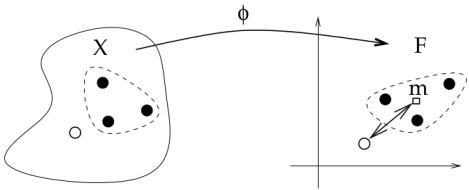
$$\lambda \tilde{\boldsymbol{v}} = \tilde{C}\tilde{\boldsymbol{v}} \tag{4}$$

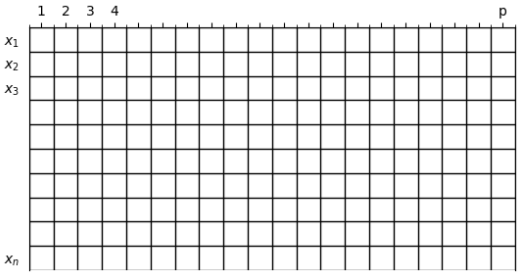
where $\tilde{\boldsymbol{v}} = \sum_{i=1}^{n} \tilde{a}_i \phi(\boldsymbol{x}_i)$

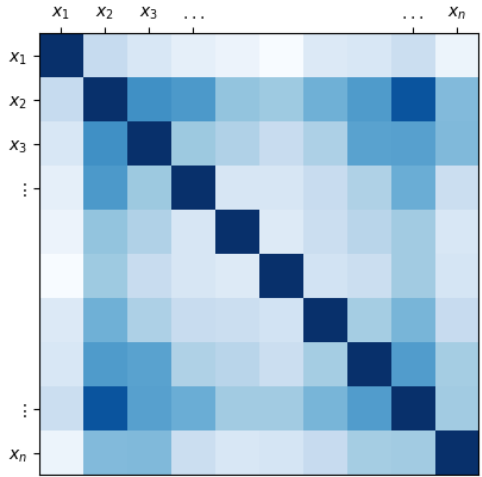
Kernel methods pose **new challenges in interpretability** as it is not easy to interpret the results in terms of the original input data.

Preimage problem

- The centroid m might have no preimage in $\mathcal{X}$.
- The distance can still be computed implicitly with the kernel trick.




1
2
3
4
p
x_1
x_2
x_3
x_n


x_1
x_2
x_3
...
...
x_n
x_1
x_2
x_3
x_n

Contents

- Why kernel PCA?
  - Reduce dimensionality
  - Non-linear method
- Improved interpretability with our method, kernel PCA Interpretable gradient, **KPCA-IG** in Briscik, Dillies, and Déjean (2023).

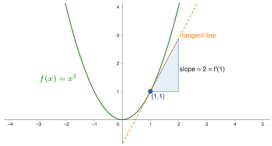

RESEARCH Open Access

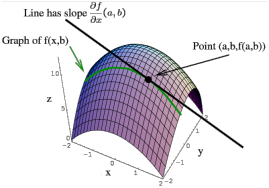
Improvement of variables interpretability in kernel PCA

Check for updates

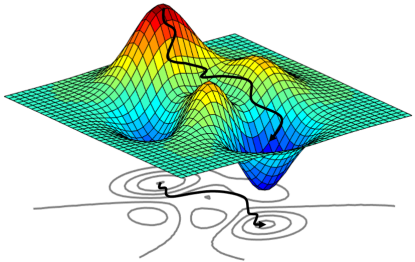
Mitja Briscik[1*], Marie-Agnès Dillies[2] and Sèbastien Déjean[1]

- **Derivative:** Measures the rate of change of a function.
- **Partial Derivative:** Represents the rate of change of a multivariate function with respect to one variable while keeping others constant.





- **Gradient Descent:** A core algorithm for training neural networks, using the gradient of the cost function to optimize weights iteratively.
- **The gradient norm** and the **direction** of the cost function play a crucial role as it contributes to the step size for each iteration, together with the learning rate.

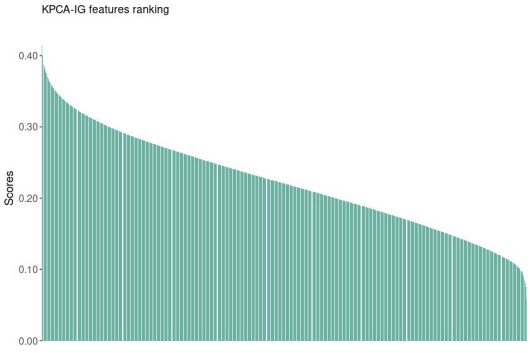


The **partial derivative of the kernel with respect to the variable of interest**, give us an indication of its the relevance for the kernel principal components.

The expression to define the effect of the variable j on the projection on the q principal components of a generic point $\boldsymbol{x}$:

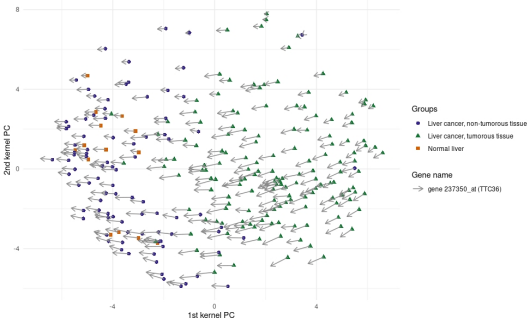
$$w^j_{1\times q} = \frac{d\varphi^j}{dt}\Big|_{t=0} = \frac{d\boldsymbol{Z}_t^T}{dt}\Big|_{t=0}\left(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\tilde{\boldsymbol{v}}, \tag{5}$$

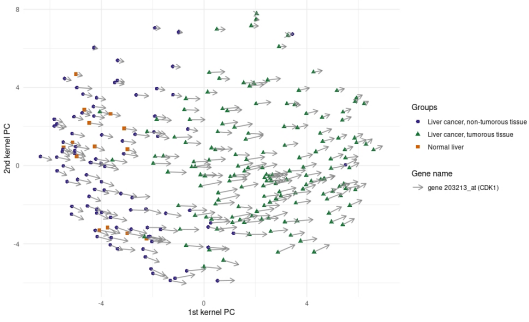
- **Compute the partial derivative** of the kernel with respect to each variable i.e. the direction of maximum variation associated with each variable for each individual
- **Average value** over all the individuals
- **Rank** of the original features
- **Display relevant variables** in the kernel principal component axes as in Reverter, Vegas, and Oller, 2014



A visual example

3 KPCA interpretability with KPCA-IG





We have 4 **main functions**:

- *kernelpca*: Kernel principal component analysis
- *kpca_igrad*: KPCA-IG: variables interpretability in kernel PCA
- *plot_kpca2D*: 2D Kernel principal analysis plot with variables representation
- *plot_kpca3D*: 3D Kernel principal analysis plot with variables representation

Package ‘kpcaIG’

June 27, 2024

Title Variables interpretability with kernel PCA

Version 1.0

Author Mitja Briscik, Mohamed Heimida, Sébastien Déjean

Maintainer Mitja Briscik <mitja.briscik@math.univ-toulouse.fr>

Description This package provides a tool for performing Kernel Principal Component Analysis (KPCA) with intepretation of the original variables. It includes functions for 2D and 3D visualization of the original variables into the kernel principal components, highlighting the contribution of specific variables using arrows.

License GPL-3

Encoding UTF-8

Imports grDevices, rgl, kernlab, ggplot2, stats, progress, viridis

NeedsCompilation no

kernelpca	*Kernel Principal Components Analysis*

Description

Kernel Principal Components Analysis, a nonlinear version of principal component analysis obtrained through the so-called kernel trick.

Usage

```
kernelpca(data, kernel = "vanilladot", kpar = list(), features = 0)
```

kpca_igrad	*KPCA-IG: variables interpretability in kernel PCA*

Description

KPCA-IG, kernel pca interpetable gradient. It is the fuction that gives the feature ranking, from the most to the least relevant variable. The ranking is obtained through the kernel's partial derivatives computation. A score, which corresponds to the score mean among the sample points, is assigned to each input feature.

Usage

```
kpca_igrad(kpca_result, dim, mean_type = "arithmetic", trim_ratio = 0.1)
```

plot_kpca2D	*2D Kernel principal analysis plot with variables representation*

Description

With this function it is possible to visualize an original variable of interest in the first two principal component. The variable is displayed as an arrow, showing its relevance in the relative position of each sample point in the kernel component space.

Usage

```
plot_kpca2D(kpca_result, target_variable, groups = NULL, scale = 100,
arrow_col = "#D3D3D3", main_title = "Kernel principal component analysis" )
```

plot_kpca3D	*3D Kernel principal analysis plot with variables representation*

Description

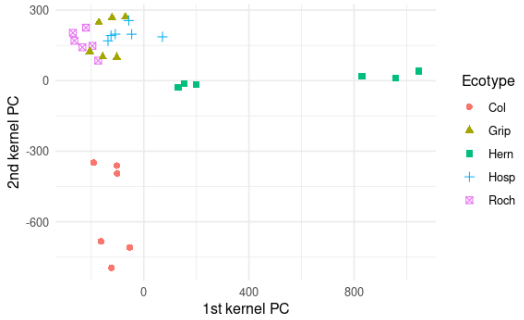
With this function it is possible to visualize an original variable of interest in the first three principal component. The variable is displayed as an arrow, showing its relevance in the relative position of each sample point in the kernel component space.

Usage

```
plot_kpca3D(kpca_result, target_variable, groups, scale=1,
type = "s", size = 3/4, arrow_col = "#999999",
angles = 12, main = NULL
)
```

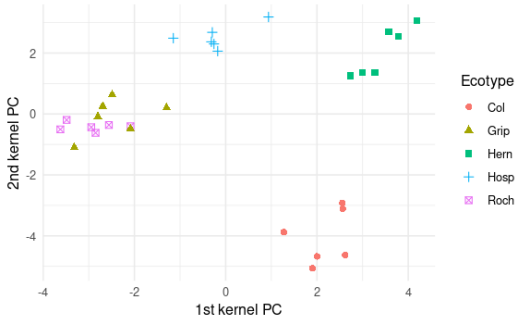
Linear PCA

```
Transcr_Stems <- scale(Transcriptomics_Stems)
kpca_linear <- kernelpca(as.matrix(Transcr_Stems), kernel = "vanilladot")
```



Kernel PCA with hyperbolic tangent kernel

```
kpca_rbf <- kernelpca(as.matrix(Transcr_Stems),
                      kernel = "rbfdot",
                      kpar = list(sigma = 0.00005)
```

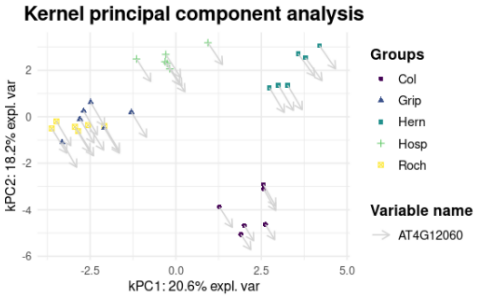


Example of usage

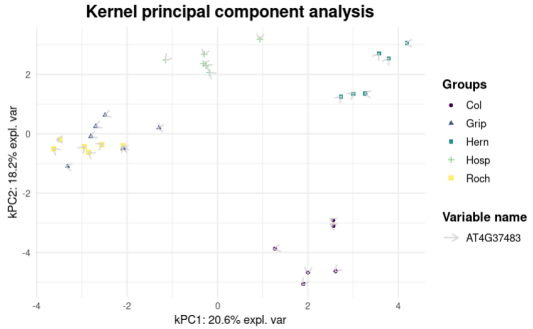
3 KPCA interpretability with KPCA-IG

```
kpcaIG_tan <- kpca_igrad(kpca_tan, dim = c(1,2))
> head(kpcaIG_tan)
  column_names means_norms    std_norms
1    AT4G12060 0.001064981 1.316034e-04
2    AT3G27420 0.001062089 1.194429e-04
3    AT5G25040 0.001059460 1.262786e-04
4    AT1G61180 0.001044166 1.308021e-04
```

```
plot_kpca2D(kpca_tan, "AT4G12060", groups = Ecotype, scale =
    1000)
```



Gene AT4G37483, the least important in the ranking of KPCA-IG.



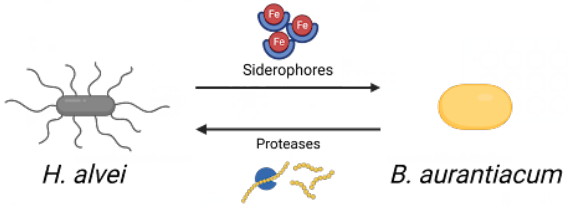
Contents

MUTUALISTIC BACTERIAL RELATIONSHIPS IN CHEESE MICROBIAL COMMUNITIES: EXPLORING INTERACTIONS THROUGH KERNEL METHODS

Mitja Briscik
Institut de Mathématiques
Université Paul Sabatier
Toulouse, 31062 FR
mitja.briscik@math.univ-toulouse.fr

Rina Mekuli
INRAE, AgroParisTech
Universitè Paris-Saclay,UMR SayFood
Palaiseau, FR
rina.mekuli@inrae.fr

- Cheese is an **iron scarce** environment
- The cross-feeding between **H. alvei** and **B. aurantiacum** allows them to access the otherwise unavailable **iron** and **nitrogen** sources.

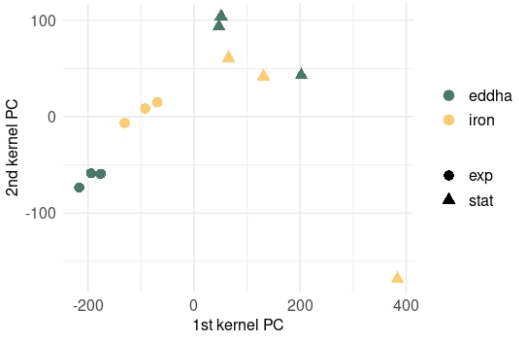


Dataset	Transcriptomics	Proteomics	Metabolomics
B. aurantiacum	4000	2583	372
Hafnia alvei	4528	2780	372

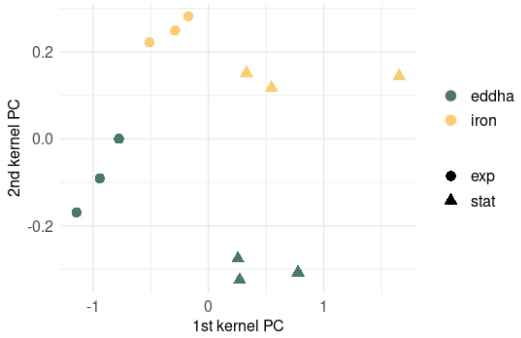
- **Condition**: Iron 6 , No-iron: 6
- **Phase**: Stationary 6 , Exponential: 6

Kernel methods are used with the goal of **identifying variables** that reveal strategies employed by ripening bacteria to **overcome iron deprivation**, ultimately leading to **microbial interactions**.

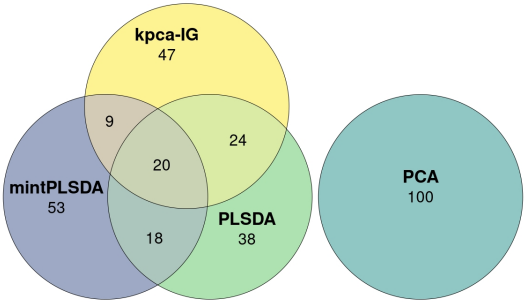
Linear PCA

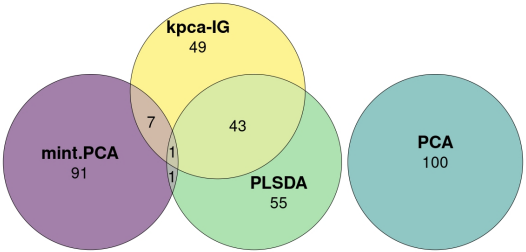


Sigmoid KPCA, $\alpha = 0.0002$ and $c = 2$

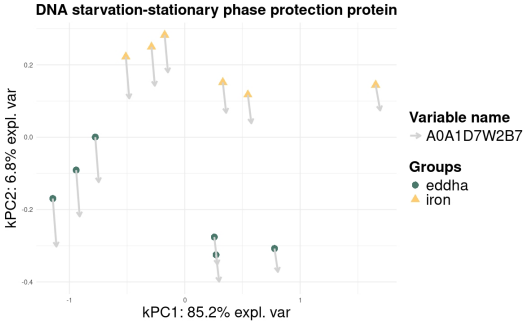


Common and unique variables - 100 first protein selected by each method

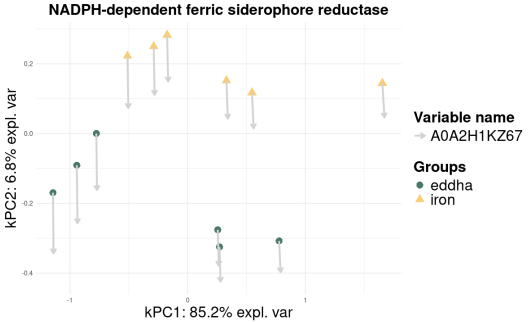




Protein A0A1D7W2B7 is known for protecting cells from oxidative stress via DNA binding and/or ferroxidase activity (Karas, Westerlaken, and Meyer, 2015). Its increased expression suggests **limited iron triggers oxidative stress**.



By **up-regulating** A0A2H1KZ67, *B. aurantiacum* likely facilitates the uptake of iron bound by *H. alvei*-produced siderophores, establishing a **mutualistic relationship in iron-limited conditions.**



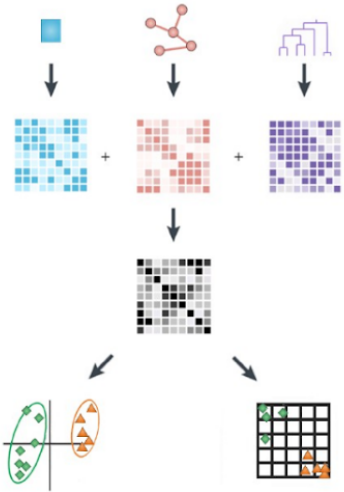
Contents

Multiple Kernel Learning (MKL)

- Analyze multiple heterogeneus sources datasets in uniform way

$$\boldsymbol{K}^* = \sum_{m=1}^{M} \beta_m \boldsymbol{K}^m \text{ subject to } \begin{cases} \beta_m \geq 0 \\ \sum_{m=1}^{M} \beta_m = 1 \end{cases}$$

$$\forall m = 1, \ldots, M$$



Mariette, J.

- Why Multiple Kernel learning?
  - Single omics analysis may not provide enough information to gain a deep understanding of a biological system (Mariette and Villa-Vialaneix, 2017)

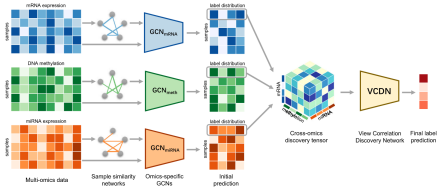

RESEARCH Open Access

Supervised multiple kernel learning approaches for multi-omics data integration

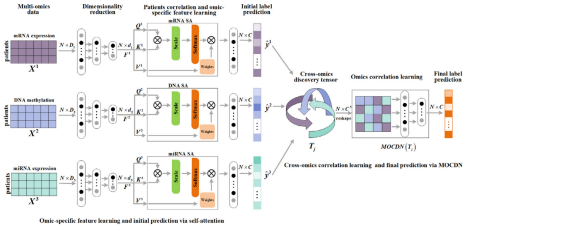
Check for updates

Mitja Briscik[1*†], Gabriele Tazza[2*†], László Vidács[2], Marie-Agnès Dillies[3] and Sébastien Déjean[1]

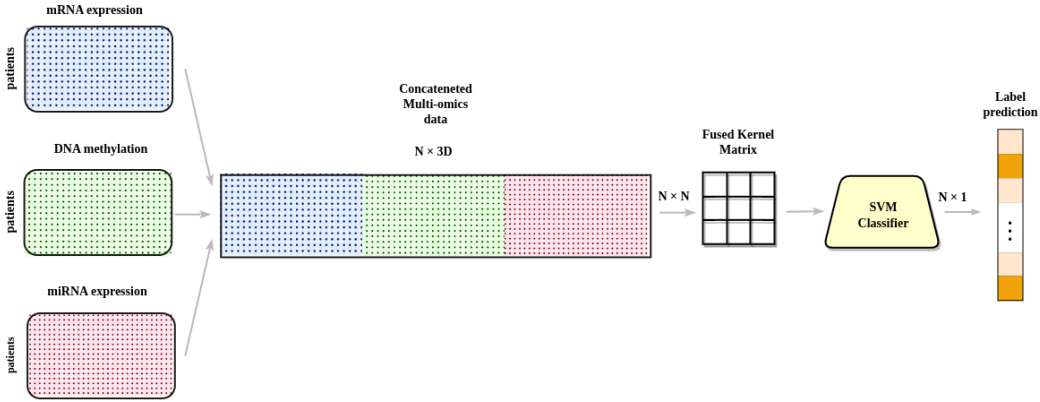
MOGONET Wang et al. (2021)



MOADLN Gong et al. (2023)

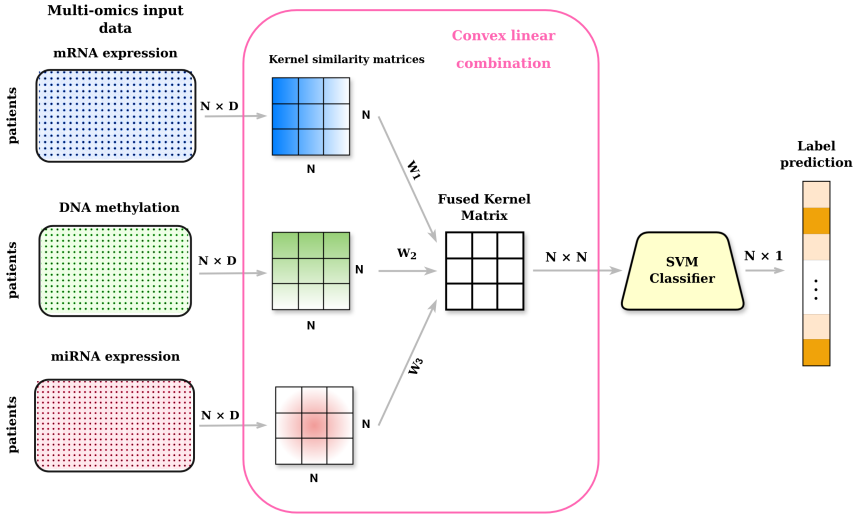


Not fair!


mRNA expression
patients
DNA methylation
patients
miRNA expression
patients
Concateneted
Multi-omics
data
N × 3D
N × N
Fused Kernel
Matrix
SVM
Classifier
N × 1
Label
prediction

SVM with mixed integration

5 Supervised multi-omics data integration with kernels



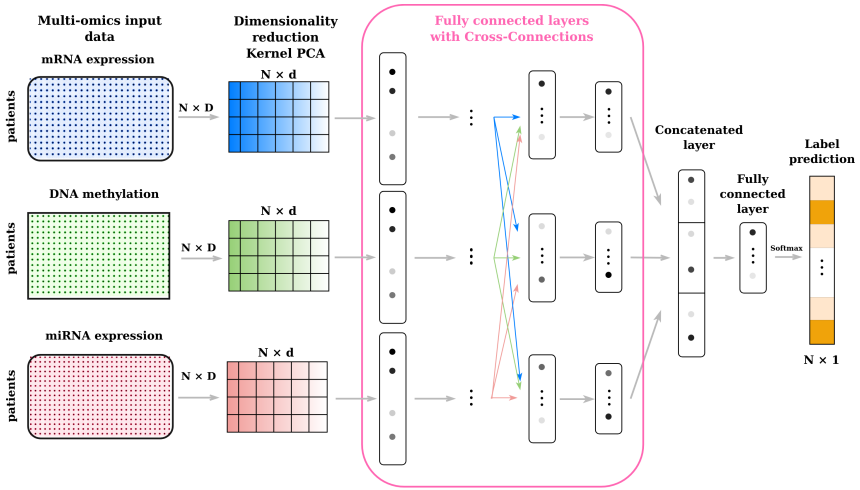

Multi-omics input data
mRNA expression
patients
N × D
Dimensionality reduction Kernel PCA
N × d
DNA methylation
patients
N × D
N × d
miRNA expression
patients
N × D
N × d
Fully connected layers with Cross-Connections
Concatenated layer
Fully connected layer
Softmax
Label prediction
N × 1

Table 3 Metrics average and standard deviation over 5 random test splits for the performance evaluation on BRCA dataset

	BRCA		
Algorithm	**ACC**	**F1_weighted**	**F1_macro**
block PLSDA	0.670 ± 0.016	0.726 ± 0.009	0.702 ± 0.011
block sPLSDA	0.668 ± 0.021	0.725 ± 0.012	0.708 ± 0.009
SVM concat	0.793 ± 0.018	0.800 ± 0.016	0.776 ± 0.017
SVM naive	0.838 ± 0.008	0.849 ± 0.008	0.828 ± 0.011
STATIS-UMKL + SVM	**0.846 ± 0.011**	**0.858 ± 0.010**	**0.837 ± 0.018**
Deep MKL (weighted sum)	0.827 ± 0.014	0.803 ± 0.015	0.831 ± 0.013
Cross-Modal Deep MKL (weighted sum)	0.829 ± 0.017	0.802 ± 0.022	0.834 ± 0.015
NN_VCDN	0.700 ± 0.018	0.692 ± 0.019	0.609 ± 0.014
Dynamics	0.826 ± 0.010	0.829 ± 0.010	0.793 ± 0.020
MOGONET	0.736 ± 0.038	0.726 ± 0.041	0.650 ± 0.053

Similar results for ROSMAP, LGG and KIPAN, where MKL-based methods, achieved competitive results outperforming claimed state-of-the-art methods.

Biomarker Discovery

A hybrid approach leveraging Deep-MKL and KPCA-IG for identifying key biomarkers.

Step 1: Selecting Relevant kernel components

- Use **Integrated Gradients** Sundararajan, Taly, and Yan (2017) to rank kernel principal components (KPCs) by their contribution to model predictions.
- Identify the most important KPCs.

Step 2: Rank original features

- Apply **KPCA-IG** to obtain a data-driven feature importance based on the selected kernel PCs.
- Use the same kernel parameters (σ) optimized in the Deep-MKL model to ensure consistency.

Identify the most important biomarkers across omics layers.

Results and Insights

- **BRCA Dataset:**
  - *mRNA Biomarkers:* GABRP, SOX10, TFF1, AGR3, SERPINB5, etc.
  - *DNA Methylation Biomarkers:* IGFBP4, RARA, NHLRC4, etc.
  - *miRNA Biomarkers:* hsa-mir-224, hsa-mir-452, hsa-mir-675, etc.
- **ROSMAP Dataset:**
  - *mRNA Biomarkers:* PREX1, CSRP1, MID1IP1, etc.
  - *DNA Methylation Biomarkers:* R3HDML, MYOD1, ALDH3B1, etc.
  - *miRNA Biomarkers:* hsa-miR-423-3p, hsa-miR-374b, hsa-miR-885-5p, etc.

Functional Insights:

- BRCA biomarkers linked to cancer progression and poor prognosis.
- ROSMAP biomarkers associated with Alzheimer's disease pathways.

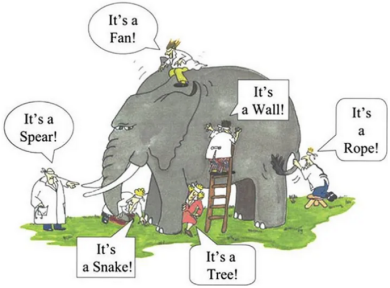
The hybrid Deep-MKL and KPCA-IG approach was found to be **effective in predicting** the disease of interest, potentially showing disease mechanisms and helping in the development of personalized treatment protocols.

Contents

Take-Home message

Multi-omics data are complex, heterogeneous, and high-dimensional, requiring advanced techniques for integration and analysis.

- **Kernel Methods:**
  - Provide a flexible, non-linear framework for data integration.
- **KPCA-IG**
  - Provides a **data-driven feature selection method** and KPCA interpretable solution.
- **Multiple Kernel Learning**
  - **MKL** showed that despite being under-utilized in multi-omics data analysis, it provides a fast and reliable solution that can compete with and **outperform more complex architectures**.
  - **Deep-MKL + KPCA-IG** successfully identified **relevant biomarkers**.



Thank you!

E-MUSE project has received funding from the European Union's Horizon 2020 research and innovation under the Marie Skłodowska-Curie grant agreement N° 956126
https://www.itn-emuse.com/, https://cordis.europa.eu/project/id/956126

References

Argelaguet, Ricard et al. (May 2021). "Computational principles and challenges in single-cell data integration". In: *Nature Biotechnology* 39.10, 1202–1215. ISSN: 1546-1696. DOI: 10.1038/s41587-021-00895-7. URL: http://dx.doi.org/10.1038/s41587-021-00895-7.

Briscik, Mitja, Marie-Agnès Dillies, and Sébastien Déjean (July 2023). "Improvement of variables interpretability in kernel PCA". In: *BMC Bioinformatics* 24.1. DOI: 10.1186/s12859-023-05404-y. URL: https://doi.org/10.1186/s12859-023-05404-y.

Gong, Ping et al. (Apr. 2023). "Multi-omics integration method based on attention deep learning network for biomedical data classification". In: *Computer Methods and Programs in Biomedicine* 231, p. 107377. DOI: 10.1016/j.cmpb.2023.107377. URL: https://doi.org/10.1016/j.cmpb.2023.107377.

References

Karas, Vlad O., Ilja Westerlaken, and Anne S. Meyer (Oct. 2015). "The DNA-Binding Protein from Starved Cells (Dps) Utilizes Dual Functions To Defend Cells against Multiple Stresses". In: *Journal of Bacteriology* 197.19. Ed. by R. L. Gourse, 3206–3215. ISSN: 1098-5530. DOI: 10.1128/jb.00475-15. URL: http://dx.doi.org/10.1128/JB.00475-15.

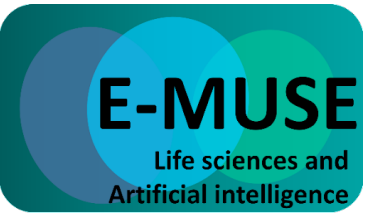
Mariette, Jérôme and Nathalie Villa-Vialaneix (Oct. 2017). "Unsupervised multiple kernel learning for heterogeneous data integration". In: *Bioinformatics* 34.6. Ed. by Jonathan Wren, pp. 1009–1015. DOI: 10.1093/bioinformatics/btx682. URL: https://doi.org/10.1093/bioinformatics/btx682.

Reverter, Ferran, Esteban Vegas, and Josep M Oller (Mar. 2014). "Kernel-PCA data integration with enhanced interpretability". In: *BMC Systems Biology* 8.S2. DOI: 10.1186/1752-0509-8-s2-s6. URL: https://doi.org/10.1186/1752-0509-8-s2-s6.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic attribution for deep networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, 3319–3328.

References

Wang, Tongxin et al. (June 2021). "MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification". In: *Nature Communications* 12.1. DOI: 10.1038/s41467-021-23774-w. URL: https://doi.org/10.1038/s41467-021-23774-w.


E-MUSE
Life sciences and
Artificial intelligence