Multi-omics data integration methods: kernel and other machine learning approaches

Nathalie Vialaneix

nathalie.vialaneix@inrae.fr
`http://www.nathalievialaneix.eu`

ML for Life Sciences
Montpellier, November 16th 2022

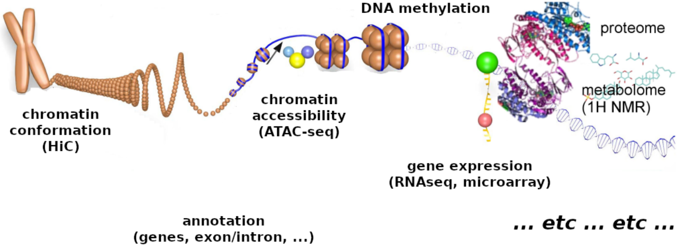
RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

Who am I?

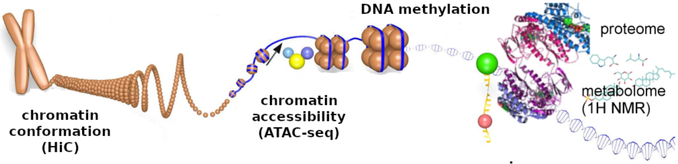
- researcher at INRAE (France)
- trained as a mathematician ⟶ statistics and machine learning for computational biology (omics)
-  mostly involved in projects on animal genomics (transcriptomics, Hi-C, ATAC-seq...)
- kernel & network methods

Collected data at genomic level are increasingly publicly available



the different levels are not always compatible

Collected data at genomic level are increasingly publicly available



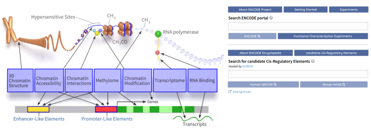
gene expression (RNAseq, microarray)

annotation (genes, exon/intron, ...)

... etc ... etc ...

the different levels are not always compatible

ENCODE: Encyclopedia of DNA Elements



TCGA

GTEx Portal

FAANG
Functional Annotation of Animal Genomes

FR-AgENCODE

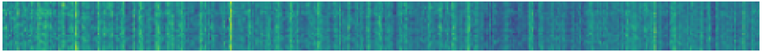
[Foissac et al., 2019]

Analysis bottlenecks

- very large dimensionality and big data (both scaling and statistical issues)

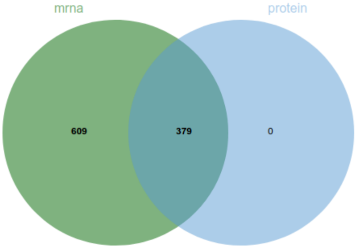
$p \sim 10^{\{3-5\}}$

$n \sim \{5 - 1000\}$



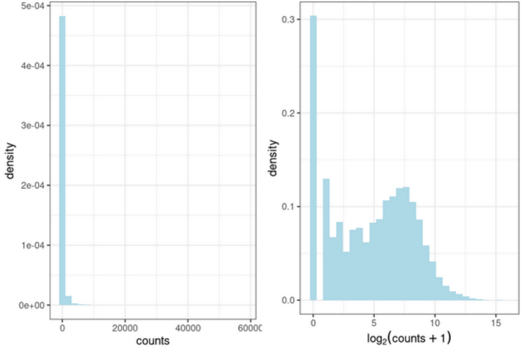
Analysis bottlenecks

- very large dimensionality and big data (both scaling and statistical issues)
- missing values and incomplete designs



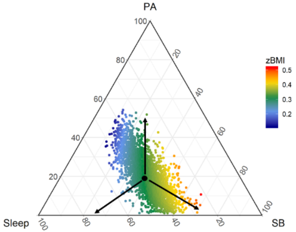
Analysis bottlenecks

- very large dimensionality and big data (both scaling and statistical issues)
- missing values and incomplete designs
- highly non Gaussian data: skewed distributions, count data, zero-inflated data, ...

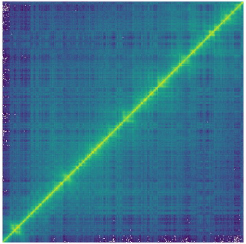


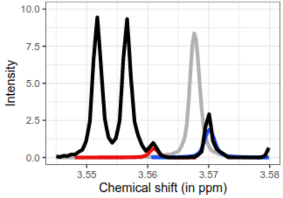
Analysis bottlenecks

- very large dimensionality and big data (both scaling and statistical issues)
- missing values and incomplete designs
- highly non Gaussian data: skewed distributions, count data, zero-inflated data, ...
- non Euclidean data: compositional data (metagenomics), similarity matrices (Hi-C), spectra (metabolomics), ...



(*)





(**)

(*) image from [Dumuid et al., 2020]

(**) image by courtesy of Gaëlle Lefort

Analysis bottlenecks

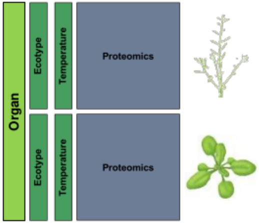
- very large dimensionality and big data (both scaling and statistical issues)
- missing values and incomplete designs
- highly non Gaussian data: skewed distributions, count data, zero-inflated data, ...
- non Euclidean data: compositional data (metagenomics), similarity matrices (Hi-C), spectra (metabolomics), ...

In addition: in plant & animal sciences, less discriminative phenotypes, interest is indirect (in breeding not directly in the individual itself), much less data with poorer quality annotation (inter-species transfer might be useful).

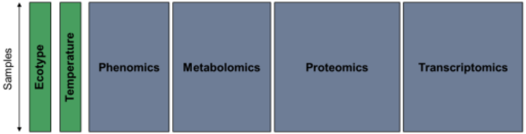
Omics data integration

Type of data to integrate

vertical:



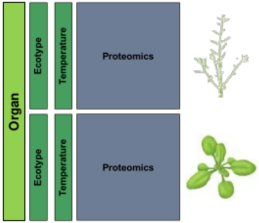
horizontal:



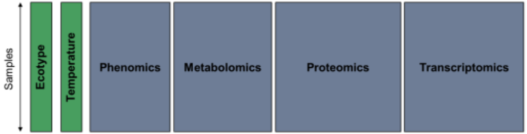
Omics data integration

Type of data to integrate

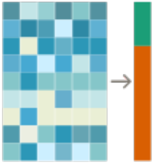
vertical:



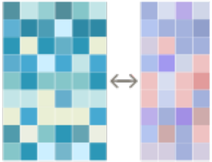
horizontal:



Type of analysis to perform



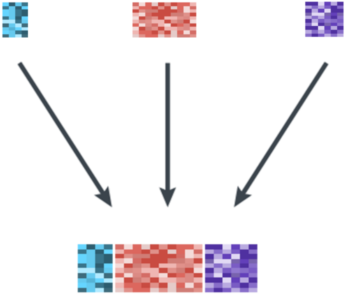
supervised:



unsupervised:

Left pictures courtesy Harold Duruflé

Multiple table analyses (CCA, MFA, PLS, STATIS, ...)



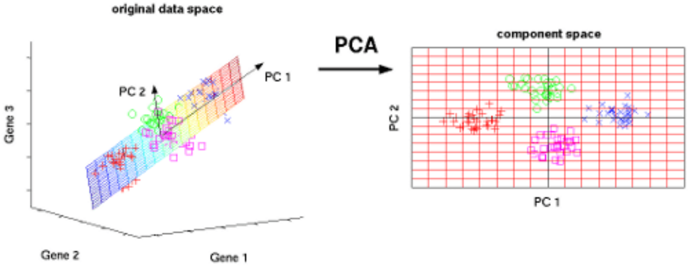
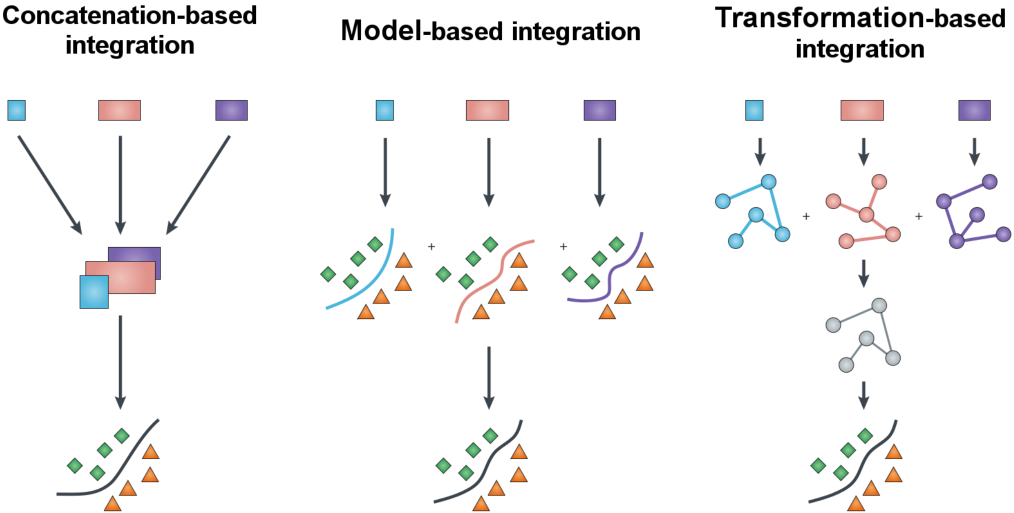


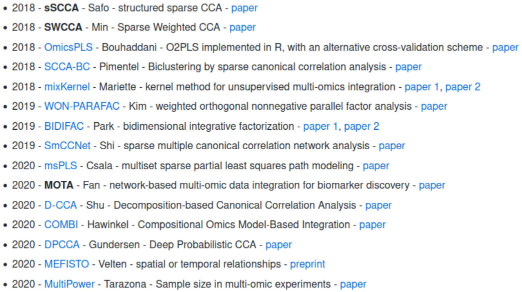
Image from `https://dimensionless.in`

Types of data integration methods [Ritchie et al., 2015]



Want to know them all?

https://github.com/mikelove/awesome-multi-omics



- 2018 - **sSCCA** - Safo - structured sparse CCA - paper
- 2018 - **SWCCA** - Min - Sparse Weighted CCA - paper
- 2018 - OmicsPLS - Bouhaddani - O2PLS implemented in R, with an alternative cross-validation scheme - paper
- 2018 - SCCA-BC - Pimentel - Biclustering by sparse canonical correlation analysis - paper
- 2018 - mixKernel - Mariette - kernel method for unsupervised multi-omics integration - paper 1, paper 2
- 2019 - WON-PARAFAC - Kim - weighted orthogonal nonnegative parallel factor analysis - paper
- 2019 - BIDIFAC - Park - bidimensional integrative factorization - paper 1, paper 2
- 2019 - SmCCNet - Shi - sparse multiple canonical correlation network analysis - paper
- 2020 - msPLS - Csala - multiset sparse partial least squares path modeling - paper
- 2020 - **MOTA** - Fan - network-based multi-omic data integration for biomarker discovery - paper
- 2020 - D-CCA - Shu - Decomposition-based Canonical Correlation Analysis - paper
- 2020 - COMBI - Hawinkel - Compositional Omics Model-Based Integration - paper
- 2020 - DPCCA - Gundersen - Deep Probabilistic CCA - paper
- 2020 - MEFISTO - Velten - spatial or temporal relationships - preprint
- 2020 - MultiPower - Tarazona - Sample size in multi-omic experiments - paper

Some specificities: can account for structure in data (network), are dedicated to a specific omic (single-cell), can account for temporal/spatial information, can include biological knowledge (mostly GO), ...

Making methods available for biologists

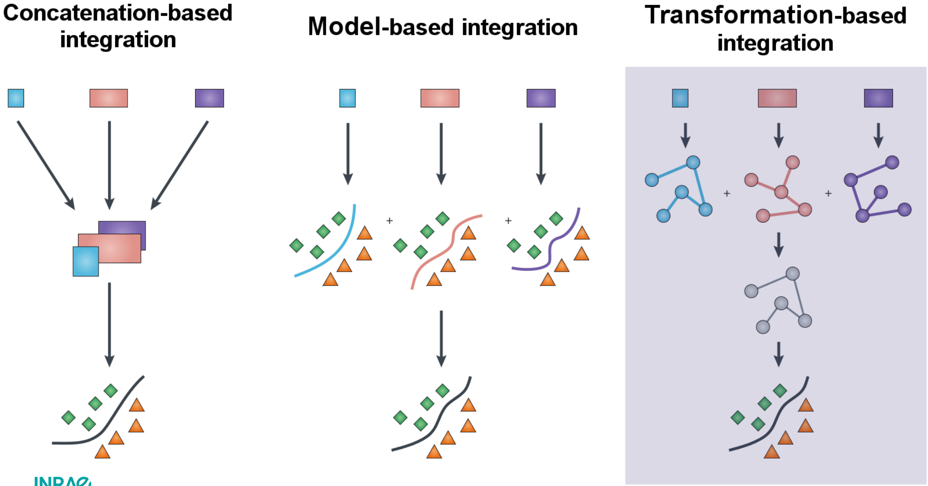
http://asterics.miat.inrae.fr

Ambition:

- easy-to-use and interactive
- helps user to know which analysis to use and how to interpret results
- integrates domain expertise
- usable online or can be installed

Scope of the rest of the talk

Unsupervised transformation based integration



Overview of the talk

Kernel methods

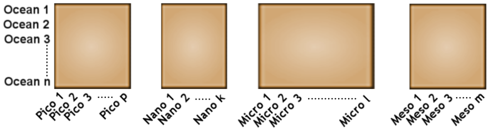
Integrating data with kernels

Improve interpretability

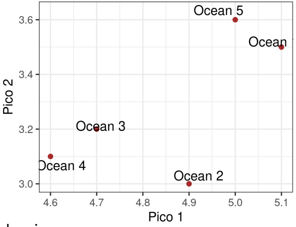
Main ideas behind kernel methods

Standard (omics) data analyses:

- data are (numeric) tables



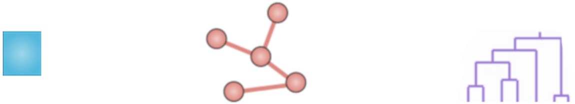
- analyses are based on operations (distances, means, ...) in the variable space



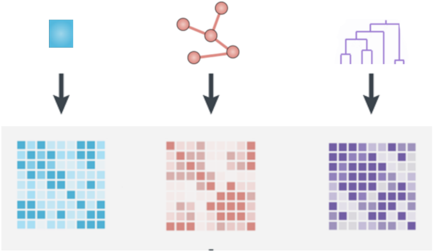
Main ideas behind kernel methods

Kernel data analyses:

- data are arbitrary



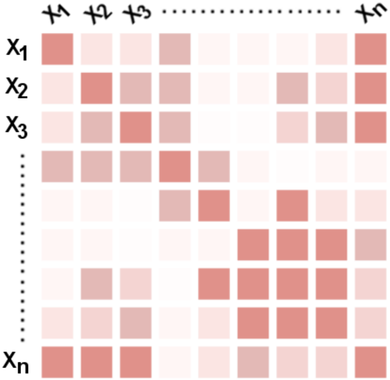
- analyses are based on transformations of data to “similarities” between samples



More formally...

n samples $(x_i)_i \in \mathcal{X}$

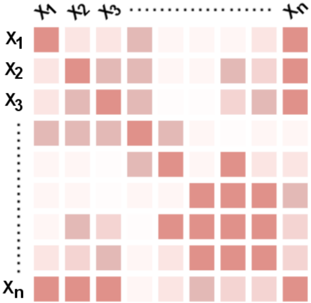
kernels: symmetric and positive definite $(n \times n)$-matrix **K** that measures a "similarity" between n entities in $\mathcal{X}$



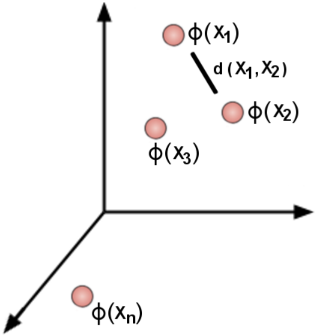
More formally...

n samples $(x_i)_i \in \mathcal{X}$

kernels: symmetric and positive definite $(n \times n)$-matrix $\mathbf{K}$ that measures a “similarity” between n entities in $\mathcal{X}$



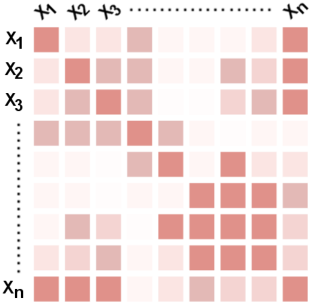
$$\mathcal{X} \xrightarrow{\Phi} \mathcal{H}$$



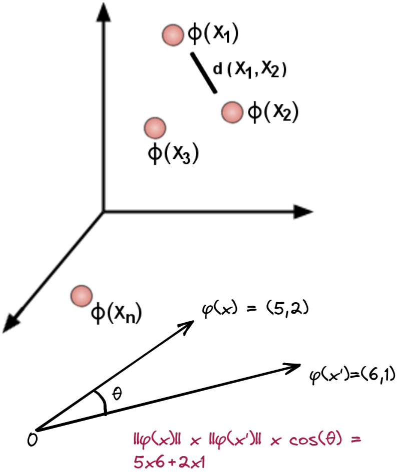
More formally...

n samples $(x_i)_i \in \mathcal{X}$

kernels: symmetric and positive definite $(n \times n)$-matrix **K** that measures a “similarity” between n entities in $\mathcal{X}$



$$\mathcal{X} \xrightarrow{\Phi} \mathcal{H}$$



$$\mathbf{K}(x, x') = \langle \phi(x), \phi(x') \rangle$$

Principles of learning from kernels

Start from any statistical method (PCA, regression, k-means clustering) and rewrite all quantities using:

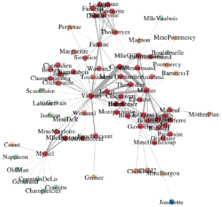
- $\mathbf{K}$ to compute distances and dot products
  dot product is: $\mathbf{K}_{ii'}$ and distance is: $\sqrt{\mathbf{K}_{ii} + \mathbf{K}_{i'i'} - 2\mathbf{K}_{ii'}}$
- (implicit) linear or convex combinations of $(\phi(x_i))_i$ to describe all unobserved elements (centers of gravity and so on...)

Kernel examples

1. $\mathbb{R}^p$ observations: Gaussian kernel $\mathbf{K}_{ii'} = e^{-\gamma \|x_i - x_{i'}\|^2}$

Kernel examples

1. $\mathbb{R}^p$ observations: Gaussian kernel $\mathbf{K}_{ii'} = e^{-\gamma \|x_i - x_{i'}\|^2}$



2. nodes of a graph: **[Kondor and Lafferty, 2002]**
3. sequence kernels (between proteins: spectrum kernel **[Jaakkola et al., 2000]** or convolution kernel **[Saigo et al., 2004]**)
4. kernel between graphs (used between metabolites based on their fragmentation trees): **[Shen et al., 2014, Brouard et al., 2016]**
5. kernel embedding phylogeny information for metagenomics **[Mariette and Villa-Vialaneix, 2018]**
6. ...

Overview of the talk

Kernel methods

Integrating data with kernels

Improve interpretability

Multiple kernel (or distance) integration

How to "optimally" combine several kernel datasets?

For kernels $\mathbf{K}^1, \ldots, \mathbf{K}^M$ obtained on the same n objects, search: $\mathbf{K}_\beta = \sum_{m=1}^M \beta_m \mathbf{K}^m$ with $\beta_m \geq 0$ and $\sum_m \beta_m = 1$

Multiple kernel (or distance) integration

How to "optimally" combine several kernel datasets?

For kernels $\mathbf{K}^1, \ldots, \mathbf{K}^M$ obtained on the same n objects, search: $\mathbf{K}_\beta = \sum_{m=1}^M \beta_m \mathbf{K}^m$ with $\beta_m \geq 0$ and $\sum_m \beta_m = 1$

- naive approach: $\mathbf{K}^* = \frac{1}{M} \sum_m \mathbf{K}^m$

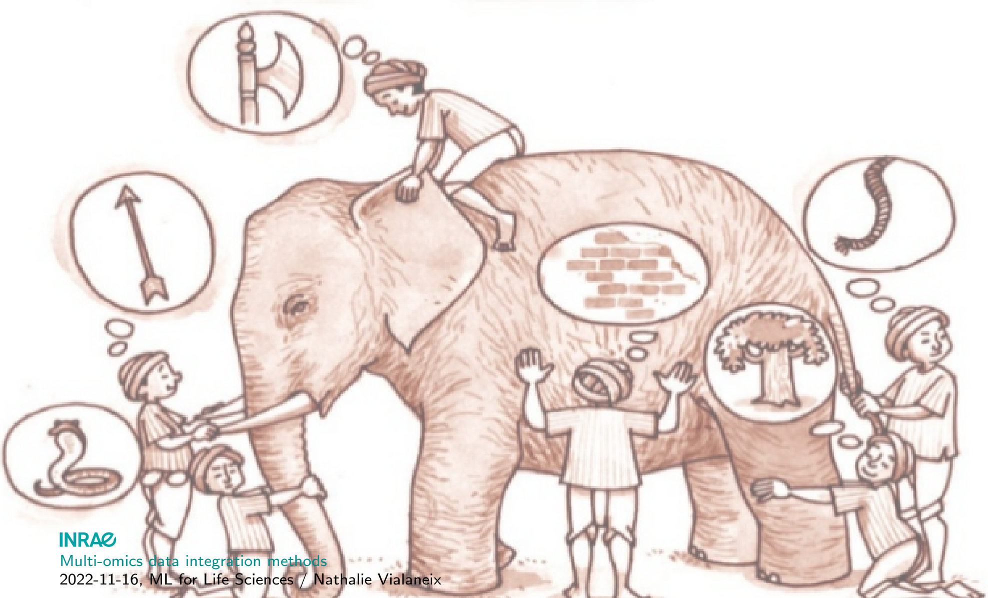
Multiple kernel (or distance) integration

How to "optimally" combine several kernel datasets?

For kernels $\mathbf{K}^1, \ldots, \mathbf{K}^M$ obtained on the same n objects, search: $\mathbf{K}_\beta = \sum_{m=1}^M \beta_m \mathbf{K}^m$ with $\beta_m \geq 0$ and $\sum_m \beta_m = 1$

- naive approach: $\mathbf{K}^* = \frac{1}{M} \sum_m \mathbf{K}^m$
- supervised framework: $\mathbf{K}^* = \sum_m \beta_m \mathbf{K}^m$ with $\beta_m \geq 0$ and $\sum_m \beta_m = 1$ with β_m chosen so as to minimize the prediction error **[Gönen and Alpaydin, 2011]**

Combining kernels in an unsupervised setting



Multiple kernel integration

Ideas of kernel consensus: find a kernel that performs a consensus of all kernels

[Mariette and Villa-Vialaneix, 2018] - R package **mixKernel**
with consensus based on:

- STATIS **[L'Hermier des Plantes, 1976, Lavit et al., 1994]**
- criterion that preserves local geometry

STATIS like framework

Similarities between kernels:

$$C_{mm'} = \frac{\langle \mathbf{K}^m, \mathbf{K}^{m'} \rangle_F}{\|\mathbf{K}^m\|_F \|\mathbf{K}^{m'}\|_F} = \frac{\mathrm{Trace}(\mathbf{K}^m \mathbf{K}^{m'})}{\sqrt{\mathrm{Trace}((\mathbf{K}^m)^2)\mathrm{Trace}((\mathbf{K}^{m'})^2)}}.$$

[Robert and Escoufier, 1976]

STATIS like framework

Similarities between kernels:

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

[Robert and Escoufier, 1976]

$$\text{maximize}_\beta \quad \sum_{m=1}^{M} \left\langle \mathbf{K}^*(\beta), \frac{\mathbf{K}^m}{\|\mathbf{K}^m\|_F} \right\rangle_F = \text{maximize}_\beta \quad \beta^\top \mathbf{C} \beta$$

$$\text{s.t. } \|\beta\|_2 = 1$$

Solution: first eigenvector of $\mathbf{C}$

TARA Oceans expedition

A smaller, faster, cheaper route to fusion energy

Sport, sex, and testosterone go to court

Making C-N bonds from nitro groups

Science

22 MAY 2015

AAAS

A world of plankton

Marine biodiversity surveyed by the schooner Tara

SHARE

RESEARCH ARTICLE

Determinants of community structure in the global plankton interactome

SHARE

RESEARCH ARTICLE

Patterns and ecological drivers of ocean viral communities

SHARE

RESEARCH ARTICLE

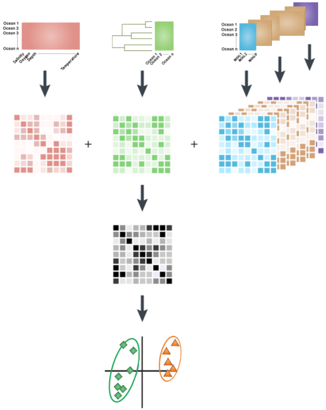
Structure and function of the global ocean microbiome

Science (May 2015) - Studies on:

- **eukaryotic plankton** diversity [de Vargas et al., 2015],
- ocean **viral communities** [Brum et al., 2015],
- global **plankton interactome** [Lima-Mendez et al., 2015],
- global **ocean microbiome** [Sunagawa et al., 2015],
- ....

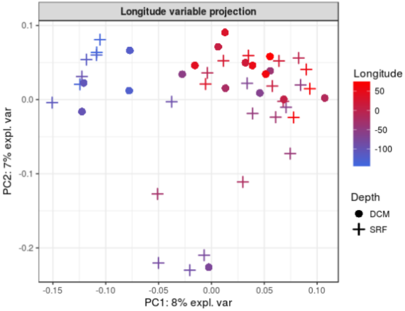
→ datasets from different types and different sources analyzed separately.

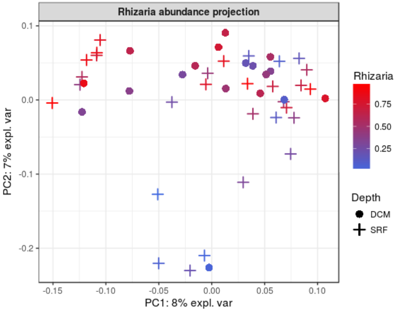
Integrating *TARA* Oceans datasets



- For all compositional datasets, include phylogenetic information (rather than CLR and alike): weighted Unifrac distance
- Perform PCA (could have been clustering, linear model, . . .) in the feature space.
  + combine with a shuffling approach to identify most influencing variables

Application to *TARA* oceans





Main facts

- Oceans typology related to **longitude**
- *Rhizaria* abundance structure the differences, especially between Arctic Oceans and Pacific Oceans

Overview of the talk

Kernel methods

Integrating data with kernels

Improve interpretability

Selecting features in kernels

Which features are important? (for numerical features only):

- label each feature j with a weight $w_j \in \{0,1\}$ (selected or not)
- new kernel: $\mathbf{K}^{\mathbf{w}}(x_i, x_{i'}) = \mathbf{K}(\mathbf{w} \cdot x_i, \mathbf{w} \cdot x_{i'})$

Selecting features in kernels

Which features are important? (for numerical features only):

- label each feature j with a weight $w_j \in \{0, 1\}$ (selected or not)
- new kernel: $\mathbf{K}^{\mathbf{w}}(x_i, x_{i'}) = \mathbf{K}(\mathbf{w} \cdot x_i, \mathbf{w} \cdot x_{i'})$
- find the best weights to optimize a quality criterion

Selecting features in kernels

Which features are important? (for numerical features only):

- label each feature j with a weight $w_j \in \{0, 1\}$ (selected or not)
- new kernel: $\mathbf{K}^{\mathbf{w}}(x_i, x_{i'}) = \mathbf{K}(\mathbf{w} \cdot x_i, \mathbf{w} \cdot x_{i'})$
- find the best weights to optimize a quality criterion

How to do it?:

- supervised framework: learn $\mathbf{w}$ to compute a kernel $\mathbf{K}_x^{\mathbf{w}}$ best to predict $Y \in \mathbb{R}$ **[Allen, 2013, Grandvalet and Canu, 2002]**

Extensions

[**Brouard et al., 2022**] and **mixKernel**

- extended to unsupervised (exploratory) learning

$$\underset{\mathbf{w}\in\{0,1\}^p}{\operatorname{argmin}} \|\mathbf{K}_x^w - \mathbf{K}_x\|_F^2 \qquad \text{s.t} \sum_{j=1}^{p} w_j \leq d$$

Continuous relaxation (non-convex and non-smooth), solved with proximal gradient descent.

Extensions

[**Brouard et al., 2022**] and **mixKernel**

- extended to unsupervised (exploratory) learning

$$\arg\min_{\mathbf{w}\in\{0,1\}^p} \|\mathbf{K}_x^w - \mathbf{K}_x\|_F^2 \qquad \text{s.t} \sum_{j=1}^{p} w_j \leq d$$

- extended to predict arbitrary outputs (kernel outputs, including multiple output regression) optimization problem

Continuous relaxation (non-convex and non-smooth), solved with proximal gradient descent.

KOKFS illustration

Used with covid19 dataset from **[Haug et al., 2020]**:

- inputs: 46 government measures (0/1 encoding)
- outputs: R time series (with a kernel based on Fréchet distances)

KOKFS illustration

Used with covid19 dataset from [**Haug et al., 2020**]:

- inputs: 46 government measures (0/1 encoding)
- outputs: R time series (with a kernel based on Fréchet distances)

More important measures:

Increase In Medical Supplies And Equipment
Border Health Check
Public Transport Restriction
The Government Provides Assistance To Vulnerable Populations
Individual Movement Restrictions
Increase Availability Of Ppe
Activate Case Notification
Border Restriction
Measures To Ensure Security Of Supply
Port And Ship Restriction
Activate Or Establish Emergency Response

Future needs for data integration

- improve interpretability of methods (integrate more biological knowledge)
- generic vs specific: omics are always evolving...
- reduce computational needs to achieve the challenge of a more sustainable research

  ```
  CO2 equivalent : 245795.0 g (~ 15.4% GIEC limit by human - 1.6tCO2e/human)
  ```

- make them more widely used by the biological community: omics data are still produced at a much faster rate than their use!

Thank you for your attention!

Questions?

References

(unofficial) Beamer template made with the help of Thomas Schiex, Matthias Zytnicki and Andreea Dreau:
`https://forgemia.inra.fr/nathalie.villa-vialaneix/bainrae`

Allen, G. I. (2013).
Automatic feature selection via weighted kernels and regularization.
Journal of Computational and Graphical Statistics, 22(2):284–299.

Brouard, C., Mariette, J., Flamary, R., and Vialaneix, N. (2022).
Feature selection for kernel methods in systems biology.
NAR Genomics and Bioinformatics, 4(1):lqac014.

Brouard, C., Shen, H., Dürkop, K., d'Alché Buc, F., Böcker, S., and Rousu, J. (2016).
Fast metabolite identification with input output kernel regression.
Bioinformatics, 32(12):i28–i36.

Brum, J., Ignacio-Espinoza, J., Roux, S., Doulcier, G., Acinas, S., Alberti, A., Chaffron, S., Cruaud, C., de Vargas, C., Gasol, J., Gorsky, G., Gregory, A., Guidi, L., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Poulos, B., Schwenck, S., Speich, S., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., *Tara* Oceans coordinators, Bork, P., Bowler, C., Sunagawa, S., Wincker, P., Karsenti, E., and Sullivan, M. (2015).
Patterns and ecological drivers of ocean viral communities.
Science, 348(6237).

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, P., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horák, A., Jaillon,

O., Lima-Mendez, G., Lukeš, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., *Tara* Oceans coordinators, Acinas, S., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M. E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., and Karsenti, E. (2015).
Eukaryotic plankton diversity in the sunlit ocean.
Science, 348(6237).

Dumuid, D., Pedišić, v., Palarea-Albaladejo, J., Martín-Fernández, J. A., Hron, K., and Olds, T. (2020).
Compositional data analysis in time-use epidemiology: what, why, how.
International Journal of Environmental Research and Public Health, 17(7):2220.

Foissac, S., Djebali, S., Munyard, K., Vialaneix, N., Rau, A., Muret, K., Esquerre, D., Zytnicki, M., Derrien, T., Bardou, P., Blanc, F., Cabau, C., Crisci, E., Dhorne-Pollet, S., Drouet, F., Faraut, T., Gonzáles, I., Goubil, A., Lacroix-Lamande, S., Laurent, F., Marthey, S., Marti-Marimon, M., Mormal-Leisenring, R., Mompart, F., Quere, P., Robelin, D., SanCristobal, M., Tosser-Klopp, G., Vincent-Naulleau, S., Fabre, S., Pinard-Van der Laan, M.-H., Klopp, C., Tixier-Boichard, M., Acloque, H., Lagarrigue, S., and Giuffra, E. (2019).
Multi-species annotation of transcriptome and chromatin structure in domesticated animals.
BMC Biology, 17:108.

Gönen, M. and Alpaydin, E. (2011).
Multiple kernel learning algorithms.
Journal of Machine Learning Research, 12:2211–2268.

Grandvalet, Y. and Canu, S. (2002).
Adaptive scaling for feature selection in SVMs.
In Becker, S., Thrun, S., and Obermayer, K., editors, *Proceedings of Advances in Neural Information Processing Systems (NIPS 2002)*, pages 569–576. MIT Press.

Haug, N., Geyrhofer, L., Londei, A., Dervic, E., Desvars-Larrive, A., Loreto, V., Pinior, B., Thurner, S., and Klimek, P. (2020).
Ranking the effectiveness of worldwide COVID-19 government interventions.
Nature Human Behaviour, 4(4):1303–1312.

Jaakkola, T., Diekhans, M., and Haussler, D. (2000).
A discriminative framework for detecting remote protein homologies.
Journal of Computational Biology, 7(1-2):95–114.

Kondor, R. and Lafferty, J. (2002).
Diffusion kernels on graphs and other discrete structures.
In Sammut, C. and Hoffmann, A., editors, *Proceedings of the 19th International Conference on Machine Learning*, pages 315–322, Sydney, Australia. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

Lavit, C., Escoufier, Y., Sabatier, R., and Traissac, P. (1994).
The ACT (STATIS method).
Computational Statistics and Data Analysis, 18(1):97–119.

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

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J., Roux, S., Vincent, F., Bittner, L., Darzi, Y., Wang, B., Audic, S., Berline, L., Bontempi, G., Cabello, A., Coppola, L., Cornejo-Castillo, F., d'Oviedo, F., de Meester, L., Ferrera, I., Garet-Delmas, M., Guidi, L., Lara, E., Pesant, S., Royo-Llonch, M., Salazar, F., Sánchez, P., Sebastian, M., Souffreau, C., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., *Tara* Oceans coordinators, Gorsky, G., Not, F., Ogata, H., Speich, S., Stemmann, L., Weissenbach, J., Wincker, P., Acinas, S., Sunagawa, S., Bork, P., Sullivan, M., Karsenti, E., Bowler, C., de Vargas, C., and Raes, J. (2015).
Determinants of community structure in the global plankton interactome.
Science, 348(6237).

Mariette, J. and Villa-Vialaneix, N. (2018).
Unsupervised multiple kernel learning for heterogeneous data integration.
Bioinformatics, 34(6):1009–1015.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015).
Methods of integrating data to uncover genotype-phenotype interactions.
Nature Reviews Genetics, 16(2):85–97.

Robert, P. and Escoufier, Y. (1976).
A unifying tool for linear multivariate statistical methods: the rv-coefficient.
Applied Statistics, 25(3):257–265.

Saigo, H., Vert, J.-P., Ueda, N., and Akutsu, T. (2004).
Protein homology detection using string alignment kernels.
Bioinformatics, 20(11):1682–1689.

Shen, H., Dührkop, K., Böcher, S., and Rousu, J. (2014).
Metabolite identification through multiple kernel learning on fragmentation trees.
Bioinformatics, 30(12):i157–i64.

Sunagawa, S., Coelho, L., Chaffron, S., Kultima, J., Labadie, K., Salazar, F., Djahanschiri, B., Zeller, G., Mende, D., Alberti, A., Cornejo-Castillo, F., Costea, P., Cruaud, C., d'Oviedo, F., Engelen, S., Ferrera, I., Gasol, J., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., *Tara* Oceans coordinators, Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S., and Bork, P. (2015).
Structure and function of the global ocean microbiome.
Science, 348(6237).

Technical details: Unsupervised Kernel Feature Selection (UKFS)

Reduced kernel for $X \in \mathbb{R}^p$

Learn weights $\mathbf{w} \in \{0,1\}^p$ such that the kernel based on $\mathbf{w} \cdot x_i$ is very similar to the kernel based on x_i

Optimization problem

$$\underset{\mathbf{w}\in\{0,1\}^p}{\operatorname{argmin}} \|\mathbf{K}_x^w - \mathbf{K}_x\|_F^2 \qquad \text{s.t} \sum_{j=1}^{p} w_j \leq d$$

Continuous relaxation (non-convex and non-smooth)

$\mathbf{w}^* := \underset{\mathbf{w}\in(\mathbb{R}^+)^p}{\operatorname{argmin}} \|\mathbf{K}_x^w - \mathbf{K}_x\|_F^2 + \lambda\|\mathbf{w}\|_1$, solved with proximal gradient descent

UKFS performances

	lapl	SPEC	MCFS	NDFS	UDFS	Autoenc.	UKFS
“Carcinom” ($n = 174$, $p = 9\ 182$)							
ACC	164.02	106.52	184.17	200.88	138.48	143.13	**206.55**
COR	28.14	30.75	29.56	27.49	30.30	33.18	**24.75**
CPU	**0.25**	2.47	11.69	6,162	99,138	$>$ 4 days	326
“Glioma” ($n = 50$, $p = 4\ 434$)							
ACC	166.31	140.72	172.78	147.77	147.50	132.76	**178.57**
COR	81.70	70.70	76.43	68.02	72.33	**45.96**	52.14
CPU	**0.02**	0.63	1.05	368	2,636	$\sim$ 12h	23.74
“Koren” ($n = 43$, $p = 980$)							
ACC	172.90	225.25	233.94	263.04	**263.48**	239.76	242.39
COR	48.18	52.34	49.94	48.48	48.69	**32.60**	47.77
CPU	**0.01**	0.07	1.11	5.88	9.70	$\sim$ 30 min	10.69

[Brouard et al., 2022]

- non redundant features
- can incorporate information on relations between variables

Technical details on kernel output feature selection (KOKFS)

Learn weights $\mathbf{w} \in \{0,1\}^p$ such that the kernel based on $\mathbf{w} \cdot x_i$ best explains the way the $(y_i)_i$ relate to each other as described by K_y:

$$\min_{h \in \mathcal{H}, \mathbf{w} \in (\mathbb{R}^+)^p} f(h, \mathbf{w}) + \lambda_1 \|h\|^2_{\mathcal{H}} + \lambda_2 \|\mathbf{w}\|_1,$$

where $f(h, \mathbf{w}) = \sum_{i=1}^n \|h(\mathbf{w} \cdot \mathbf{x}_i) - \psi(y_i)\|^2_{\mathcal{F}_y}$ and h of the following form: $h(x_i) = V\phi(\mathbf{x}_i)$.

This optimization problem is solved using an iterative algorithm alterning optimization of **w** (similar to unsupervised framework) and optimization of h (using kernel trick).

Back

General summary of KOKFS

$$\begin{tikzcd}
\mathcal{F}_x \arrow[r, "V"] & \mathcal{F}_y \\
\mathcal{X} = \mathbb{R}^p \arrow[u, "\phi_{\mathbf{w}}"] \arrow[ur, red, "h"] & \mathcal{Y} \arrow[u, "\psi"']
\end{tikzcd}$$