Optimal transport-based machine learning to match specific expression patterns in omics data

Olivier Bouaziz

Joint work with:

T. T. Y. Nguyen, W. Harchaoui, L. Mégret, C. Mendoza, C. Neri and A. Chambaz

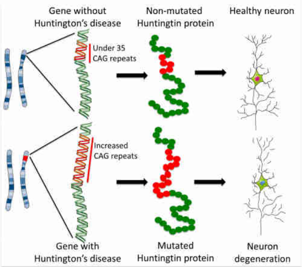
Visite HCERES - MAP5

January 10th, 2024

Huntington's disease (HD)

- HD is a progressive brain disorder that causes uncontrolled movements, emotional problems, and loss of thinking ability (cognition).
- HD is caused by the HTT gene's mutation.
  - In normal people, the HTT gene contains a triple CAG repeat about 10-35 times.
  - In people with HD, this repeat goes on for 36 or more times.

  Onset of disease occurs earlier and deterioration is faster with higher number of CAG repeat.
- HD leads to neuronal cell death.



Gene without Huntington's disease
Non-mutated Huntingtin protein
Healthy neuron
Under 35 CAG repeats
Increased CAG repeats
Gene with Huntington's disease
Mutated Huntingtin protein
Neuron degeneration

Figure: CAG repeat expansions in HD. Source: California's Stem Cell Agency.

Micro RNAs (miRNAs) and messenger RNAs (mRNAs)

- mRNAs are necessary for translating the genetic information into proteins.
- miRNAs are able to turn off genes by inactivating mRNAs.
- A miRNA is complementary to a part of one or more mRNAs, that promotes cleavage or destroy them.
- **The miRNAs and their target-mRNAs have a many-to-many mirroring relationship**

  → We will use this property.

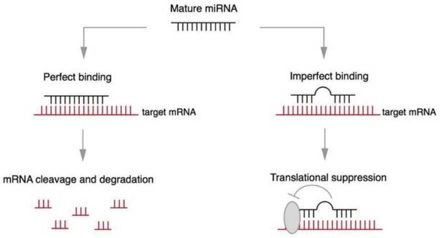


Figure: Mechanism of miRNA action. MiRNA can bind to specific regions of target mRNA transcripts and destabilizes the target transcript and/or blocks its translation. Source: [11].

Experiment and data

- *Striatum* of knock-in HD mice.
- Intervention on *polyQ* (CAG) length, one of $\{20, 80, 92, 111, 140, 175\}$.
- *Time* of evaluation of miRNA and mRNA expressions (log-fold change), either 2, 6 or 10 months.
- Results in $M = 13,616$ (mRNA) and $N = 1,143$ (miRNA) **profiles** (data points) in $\mathbb{R}^{15}$.

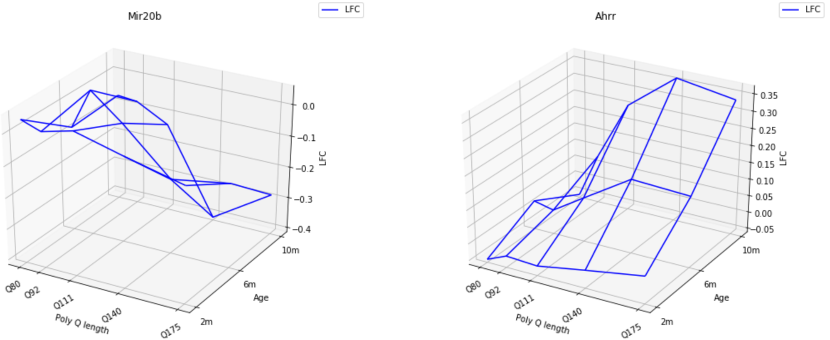


Figure: Left: profile y_n of a miRNA (Mir20b). Right: profile x_m of a mRNA (Ahrr). It is believed that Mir20b targets Ahrr.

Objective

Finding couples (miRNA, mRNA) that "collaborate"

Based on the profiles $\{x_1, \ldots, x_M\}$ (mRNA profiles) and $\{y_1, \ldots, y_N\}$ (miRNA profiles), we wish to identify collections $\{(x_m, y_n) : (m, n) \in \mathcal{S}\}$ gathering mRNAs and miRNAs that "collaborate".

- An *ideal* match between a mRNA and a miRNA would consist of two profiles that display a *perfect* mirroring relationship: $y_n = -x_m$.
- We will relax this very strong biological hypothesis and consider loosened relationships $y \approx \theta(x)$ for a transformation $\theta \in \Theta$, where Θ is a parametric set containing $-\text{id}$.

- Illustration: profiles of two mRNA and miRNA which are believed to collaborate

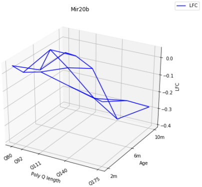


Figure: Profile of Mir20b (miRNA)

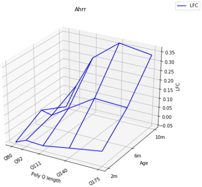


Figure: Profile of Ahrr (mRNA)

Proposal

- We develop a procedure called WTOT-matching to find collections $\{(x_m, y_n) : (m, n) \in S\}$ of mRNAs and miRNAs that "collaborate".
- The procedure unfolds in two steps:

  WTOT-...: consists in constructing a *similarity matrix* between mRNAs and miRNAs

  - we define the *similarity matrix* as an *optimal coupling matrix*;
  - we operationalize the search of mirroring relationships.

  ...-matching: consists in deriving several sets of matched elements from the similarity matrix.

Modicum of optimal transport (1/2)

- Let $X := \{x_1, \ldots, x_M\} \subset \mathbb{R}^d$ and $Y := \{y_1, \ldots, y_N\} \subset \mathbb{R}^d$ be two data sets.
- For any $\omega \in \Omega_M := \{o \in (\mathbb{R}_+)^M : \text{ st } \|o\|_1 = 1\}$ and $\omega' \in \Omega_N$, define

$$\mu_X^\omega = \sum_{m \in [\![M]\!]} \omega_m \delta_{x_m}, \qquad \nu_Y^{\omega'} = \sum_{n \in [\![N]\!]} \omega'_n \delta_{y_n}.$$

  - The measures μ_X^ω and $\mu_Y^{\omega'}$ represent the two data sets.
  - Each x_m is given a weight ω_m.
  - Each y_n is given a weight ω'_n.

- The optimal transport (OT) matrix is defined as any element of

$$\underset{P \in \Pi(\omega, \omega')}{\arg\min} \langle C_{X,Y}, P \rangle_F,$$

  where

  - $\Pi(\omega, \omega')$ is the set of $P \in (\mathbb{R}_+)^{M \times N}$ such that $P\,\mathbf{1}_N = \omega$ and $P^\top\,\mathbf{1}_M = \omega'$;
  - $C_{X,Y} \in \mathbb{R}^{M \times N}$ is a cost matrix given by $(C_{X,Y})_{mn} := c(x_m, y_n)$ for some cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$;
  - $\langle C_{X,Y}, P \rangle_F := \sum_{(m,n) \in [\![M]\!] \times [\![N]\!]} (C_{X,Y})_{mn} P_{mn}$

- Computing the $\arg\min$ is difficult and slow (and unicity is not guaranteed).

Modicum of optimal transport (2/2)

- Focus on entropic-regularized OT: for any $\gamma > 0$,

$$\mathcal{W}_\gamma\left(\mu_X^\omega, \nu_Y^{\omega'}\right) = \min_{P \in \Pi(\omega,\omega')} \left\{\langle C_{X,Y}, P\rangle_F - \gamma E(P)\right\}$$

where $E(P) = -\sum_{(m,n)\in [\![M]\!]\times[\![N]\!]} P_{mn}(\log P_{mn} - 1)$. Gain?

  - unique minimizer;
  - computing the $\arg\min$ is much easier (Sinkhorn's algorithm).

- Introduce the Sinkhorn loss:

$$\bar{\mathcal{W}}_\gamma(\mu_X^\omega, \nu_Y^{\omega'}) := 2\mathcal{W}_\gamma(\mu_X^\omega, \nu_Y^{\omega'}) - \mathcal{W}_\gamma(\mu_X^\omega, \mu_X^\omega) - \mathcal{W}_\gamma(\nu_Y^{\omega'}, \nu_Y^{\omega'})$$

Gain?

  - non-negative, symmetric, convex;
  - metrizes convergence of measures;
  - unbiased gradient estimates;
  - interpolates between OT (its nice geometry) and Maximum Mean Discrepancy (its favorable high-dimensional sample complexity + sensitivity to differences in both location and shape of distributions).

WTOT-matching: WTOT-... (1/4)

- Introduce

$$\Theta := \left\{\theta : \mathbb{R}^d \to \mathbb{R}^d, x \mapsto \theta(x) = \theta_1 x + \theta_2, \theta_1 \in T_1 \subset \mathbb{R}^{d\times d}, \theta_2 \in \mathbb{R}^d\right\},$$

  where

  - the matrices θ_1 are constrained;
  - in particular, their diagonals are made of negative values ($\sim$ mirroring relationship);
  - $-\mathrm{id} \in \Theta$.

- For all $\omega \in \Omega_M, \theta \in \Theta$, define

$$\mu^{\omega}_{\theta(X)} = \sum_{m\in[\![M]\!]} \omega_m \delta_{\theta(X_m)}, \qquad \nu_Y = \frac{1}{N}\sum_{n\in[\![N]\!]} \delta_{Y_n}.$$

- Our master program is

$$\min_{\omega\in\Omega}\min_{\theta\in\Theta} \bar{\mathcal{W}}_\gamma\left(\mu^{\omega}_{\theta(X)}, \nu_Y\right), \qquad (\square)$$

  where we are interested in the minimizer$(\hat{\omega}, \hat{\theta})$ *and* in the optimal matrix $\hat{P} \in \Pi(\hat{\omega}, N^{-1}\,\mathbf{1}_N)$ solving

$$\min_{P\in\Pi(\hat{\omega},N^{-1}\,\mathbf{1}_N)} \left\{\langle C_{\hat{\theta}(X),Y}, P\rangle_F - \gamma E(P)\right\}.$$

WTOT-matching: WTOT-... (2/4)

- We propose to solve (□) by iteratively updating ω and then θ.
- Given a kernel φ (standard normal density):
  - sample $\theta^{(0)}$ in Θ;
  - iteratively for $0 \leq \tau < T$,
    1. define $\omega^{(\tau)} \in \Omega_M$ such that $\omega_m^{(\tau)} \propto \nu_Y \varphi\left(\frac{\cdot - \theta^{(\tau)}(x_m)}{h}\right)$ (all $m \in [\![M]\!]$);
    2. solve $\theta^{(\tau+1)} \in \arg\min_{\theta \in \Theta} \bar{\mathcal{W}}_\gamma\left(\mu^{\omega^{(\tau)}}_{\theta(X)}, \nu_Y\right)$.
- Then we retrieve the corresponding OT matrix $\hat{P}$ that solves

$$\mathcal{W}_\gamma\left(\mu^{\omega^{(T)}}_{\theta^{(T)}(X)}, \nu_Y\right) = \min_{P \in \Pi(\omega^{(T)}, N^{-1}1_N)} \left\{\langle C_{\theta^{(T)}(X), Y}, P\rangle_F - \gamma E(P)\right\}.$$

- Comments:
  - use of mini-batches in step 2;
  - $\theta^{(T)} : \mathbb{R}^d \to \mathbb{R}^d$ models to relax the mirroring relationships;
  - $\hat{P}_{mn}$ can be interpreted as a similarity between x_m and y_n;
  - $\omega^{(T)}$: weights to operationalize the many-to-many relationships.

WTOT-matching: ...-matching (3/4)

- Fix two integers $k, k' \geq 1$, let $\hat{\tau}$ be the quantile of order q of all the entries of $\hat{P}$.
- For every $m \in [\![M]\!]$ and $n \in [\![N]\!]$

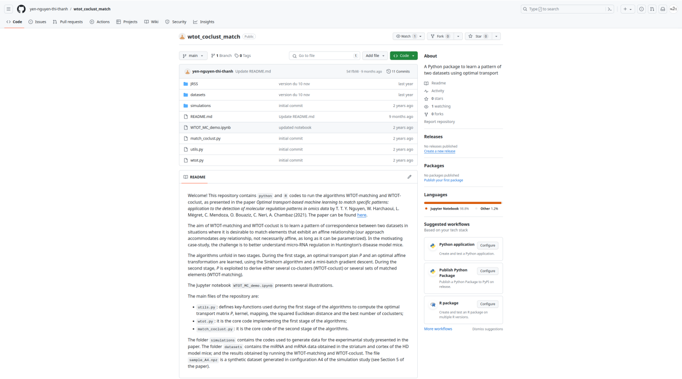
$$\mathcal{N}_m^0 := \Big\{ n \in [\![N]\!] : \hat{P}_{mn} \in \{\hat{P}_{m(1)}, \dots, \hat{P}_{m(k)}\} \text{ and } \hat{P}_{mn} \geq \hat{\tau} \Big\},$$
$$\mathcal{M}_n^0 := \Big\{ m \in [\![M]\!] : \hat{P}_{mn} \in \{\hat{P}_{(1)n}, \dots, \hat{P}_{(k')n}\} \text{ and } \hat{P}_{mn} \geq \hat{\tau} \Big\}.$$

- Define the most relevant matches

$$\mathcal{R} := \Big\{ (m, n) \in [\![M]\!] \times [\![N]\!] : n \in \mathcal{N}_m^0 \text{ and } m \in \mathcal{M}_n^0 \Big\}.$$

WTOT-matching: code (4/4)

- Code written in `python` and available on this webpage.



yen-nguyen-thi-thanh Update README.md 541fb98 · 9 months ago 11 Commits

JRSS	version du 10 nov	last year
datasets	version du 10 nov	last year
simulations	initial commit	2 years ago
README.md	Update README.md	9 months ago
WTOT_MC_demo.ipynb	updated notebook	2 years ago
match_coclust.py	initial commit	2 years ago
utils.py	initial commit	2 years ago
wtot.py	initial commit	2 years ago

README

Welcome! This repository contains `python` and `R` codes to run the algorithms WTOT-matching and WTOT-coclust, as presented in the paper *Optimal transport-based machine learning to match specific patterns: application to the detection of molecular regulation patterns in omics data* by T. T. Y. Nguyen, W. Harchaoui, L. Mégret, C. Mendoza, O. Bouaziz, C. Neri, A. Chambaz (2021). The paper can be found here.

The aim of WTOT-matching and WTOT-coclust is to learn a pattern of correspondence between two datasets in situations where it is desirable to match elements that exhibit an affine relationship (our approach accommodates *any* relationship, not necessarily affine, as long as it can be parametrized). In the motivating case-study, the challenge is to better understand micro-RNA regulation in Huntington's disease model mice.

The algorithms unfold in two stages. During the first stage, an optimal transport plan P and an optimal affine transformation are learned, using the Sinkhorn algorithm and a mini-batch gradient descent. During the second stage, P is exploited to derive either several co-clusters (WTOT-coclust) or several sets of matched elements (WTOT-matching).

The Jupyter notebook `WTOT_MC_demo.ipynb` presents several illustrations.

The main files of the repository are:

- `utils.py` : defines key-functions used during the first stage of the algorithms to compute the optimal transport matrix P, kernel, mapping, the squared Euclidean distance and the best number of coclusters;
- `wtot.py` : it is the core code implementing the first stage of the algorithms;
- `match_coclust.py` : it is the core code of the second stage of the algorithms.

The folder `simulations` contains the codes used to generate data for the experimantal study presented in the paper. The folder `datasets` contains the miRNA and mRNA data obtained in the striatum and cortex of the HD model mice; and the results obtained by running the WTOT-matching and WTOT-coclust. The file `sample_A4.npz` is a synthetic dataset generated in configuration A4 of the simulation study (see Section 5 of the paper).

About

A Python package to learn a pattern of two datasets using optimal transport

Readme
Activity
0 stars
1 watching
0 forks
Report repository

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

Jupyter Notebook 98.8% Other 1.2%

Suggested workflows
Based on your tech stack

Python application
Create and test a Python application.

Publish Python Package
Publish a Python Package to PyPI on release.

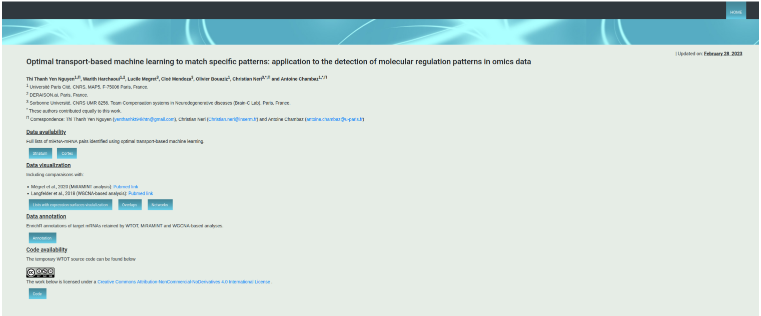
R package
Create and test an R package on multiple R versions.

More workflows Dismiss suggestions

- A tutorial is made available to show how simple it is to run the code.
- We adapt the Sinkhorn algorithm implemented by Aude Genevay and available here.
- The stochastic gradient descents relies on the machine learning framework `pytorch`.

Real data application (1/4)

- We choose $k = k' = 10$, $q = 90\%$.
- Some facts:
  - we obtain 4234 non-empty $\mathcal{N}_m$s and 1043 non-empty $\mathcal{M}_n$s;
  - $\frac{\sum_{m\in[\![M]\!]} \mathrm{card}(\mathcal{N}_m)}{\{m\in[\![M]\!]:\mathcal{N}_m\neq\emptyset\}} \approx 1.82$, $\frac{\sum_{n\in[\![N]\!]} \mathrm{card}(\mathcal{M}_n)}{\{n\in[\![N]\!]:\mathcal{M}_n\neq\emptyset\}} \approx 6.04$.
- Our findings and their analysis are shared on this website.



| Updated on: February 28 2023

Optimal transport-based machine learning to match specific patterns: application to the detection of molecular regulation patterns in omics data

Thi Thanh Yen Nguyen[1,Π], Warith Harchaoui[1,2], Lucile Megret[3], Cloé Mendoza[3], Olivier Bouaziz[1], Christian Neri[3,*,Π] and Antoine Chambaz[1,*,Π]

[1] Université Paris Cité, CNRS, MAP5, F-75006 Paris, France.

[2] DERAISON.ai, Paris, France.

[3] Sorbonne Université, CNRS UMR 8256, Team Compensation systems in Neurodegenerative diseases (Brain-C Lab), Paris, France.

[*] These authors contributed equally to this work.

[Π] Correspondence: Thi Thanh Yen Nguyen (yenthanhkt94khtn@gmail.com), Christian Neri (Christian.neri@inserm.fr) and Antoine Chambaz (antoine.chambaz@u-paris.fr)

Data availability

Full lists of miRNA-mRNA pairs identified using optimal transport-based machine learning.

Data visualization

Including comparaisons with:

- Mégret et al., 2020 (MiRAMINT analysis): Pubmed link
- Langfelder et al., 2018 (WGCNA-based analysis): Pubmed link

Data annotation

EnrichR annotations of target mRNAs retained by WTOT, MiRAMINT and WGCNA-based analyses.

Code availability

The temporary WTOT source code can be found below

Real data application: example of “monotonic” profiles (2/4)

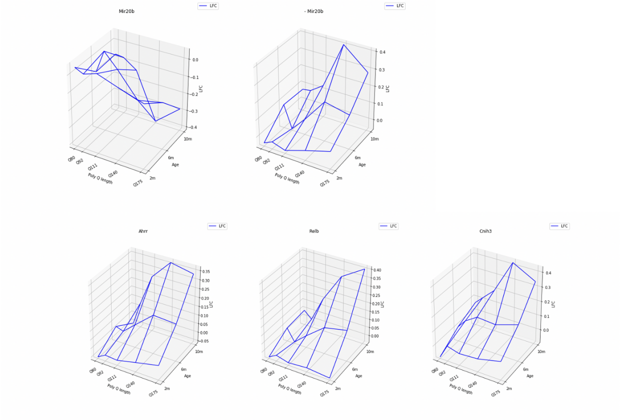


Figure: **Top:** profile y_n of Mir20b (left) and $-y_n$ (right). **Bottom:** profiles x_m ($m \in \mathcal{M}_n$) of the matched mRNAs Ahrr, Relb and Cnih3.

Real data application: example of “peaked” profiles (3/4)

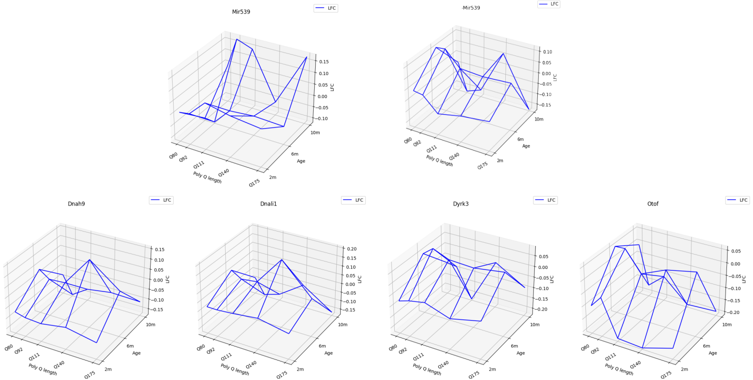


Figure: **Top:** profile y_n of Mir539 (left) and $-y_n$ (right). **Bottom:** profiles x_m ($m \in \mathcal{M}_n$) of the matched mRNAs Dnah9, Dnali1, Dyrk3, Otof.

Real data application: biological analysis of the results (4/4)

- A biological analysis is conducted to identify the more relevant pairs.
- A pair (x, y) is retained if and only if the mRNA whose profile is x and the miRNA whose profile is y are both among the 27,355 mRNAs and 1,478 miRNAs appearing in the TargetScan [5], MicroCosm [1] and miRDB [3] databases.
- The enrichment analysis reveals that the matchings output by WTOT-matching are
  1. primarily annotated for *extracellular matrix organization*, which relates to cell identity (due to the matchings labeled as neither peaked nor monotonic);
  2. secondarily annotated for *mitigation of host antiviral defense response* (due to the matchings labeled as monotonic), and for *conventional motile cilium* (due to the matchings labeled as peaked).

Real data application: biological analysis of the results (4/4)

- A biological analysis is conducted to identify the more relevant pairs.
- A pair (x, y) is retained if and only if the mRNA whose profile is x and the miRNA whose profile is y are both among the 27,355 mRNAs and 1,478 miRNAs appearing in the TargetScan [5], MicroCosm [1] and miRDB [3] databases.
- The enrichment analysis reveals that the matchings output by WTOT-matching are
  1. primarily annotated for *extracellular matrix organization*, which relates to cell identity (due to the matchings labeled as neither peaked nor monotonic);
  2. secondarily annotated for *mitigation of host antiviral defense response* (due to the matchings labeled as monotonic), and for *conventional motile cilium* (due to the matchings labeled as peaked).

Thanks for your attention!

References I

[1] D. Betel, A. Koppal, P. Agius, C. Sander, and C. Leslie. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biology*, 11:R90, 2010.

[2] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[3] J. Ding, X. Li, and H. Hu. TarPmiR: a new approach for microRNA target site prediction. *BMC Bioinformatics*, 32:2768–2775, 2016.

[4] P. Langfelder, F. Gao, N. Wang, D. Howland, S. Kwak, T. Vogt, J. Aaronson, J. Rosinski, G. Coppola, S. Horvath, and X. Yang. MicroRNA signatures of endogenous Huntingtin CAG repeat expansion in mice. *PloS One*, 13(1), 2018.

[5] Benjamin P. Lewis, Christopher B. Burge, and David P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, 2005.

[6] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

[7] S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

References II

[8] L. Mégret, S. Sasidharan Nair, J. Dancourt, J. Aaronson, J. Rosinski, and C. Neri. Combining feature selection and shape analysis uncovers precise rules for miRNA regulation in Huntington's disease mice. *BMC Bioinformatics*, 21(1):75, 2020.

[9] P. Ochs, T. Brox, and T. Pock. iPiano: inertial proximal algorithm for strongly convex optimization. *J. Math. Imaging Vision*, 53(2):171–181, 2015.

[10] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[11] A. L. Teixeira, F. Dias, M. Gomes, M. Fernandes, and R. Medeiros. Circulating biomarkers in renal cell carcinoma: the link between microRNAs and extracellular vesicles, where are we now? *Journal of Kidney Cancer and VHL*, 1:84–98, 12 2014.