Improving distance measures between genomic tracks with mutual proximity

Thomas Haschka, Jean Baptiste Morlot, Leopold Carron, Julien Mozziconacci

PAPER

Improving distance measures between genomic tracks with mutual proximity

Thomas Haschka,[1,*] Jean Baptiste Morlot,[2] Leopold Carron[2] and Julien Mozziconacci[1,2,*]

[1]Structure et Instabilité des Génomes (StrInG), Museum National d'Histoire Naturelle, INSERM, CNRS, Alliance Sorbonne Université, 75005, Paris, France and [2]Physics Department, Theoretical Physics of Condensed Matter Lab, Sorbonne Université, 75005, Paris, France
[*]Corresponding authors: haschka@gmail.com, julien.mozziconacci@mnhn.fr

Abstract

An increasing number of genomic tracks such as DNA methylation, histone modifications or transcriptomes are being produced to annotate genomes with functional states. The comparison of such high dimensional vectors obtained under various experimental conditions requires the use of a distance or dissimilarity measure. Pearson, Cosine and L_p-norm distances are commonly used for both count and binary vectors. In this article we highlight how enhancement methods such as the contrast increasing *mutual proximity* or *local scaling* improves common distance measures. We present a systematic approach to evaluate the performance of such enhanced distance measures in terms of separability of groups of experimental replicates to outline their effect. We show that the *mutual proximity* applied on the various distance measures drastically increases performance. Depending on the type of epigenetic experiment, *mutual proximity* coupled together with Pearson, Cosine, L_1, Yule or Jaccard distances, proves to be highly efficient in discriminating epigenomic profiles.

Key words: ChIP-seq, RNA-seq, DNA methylation, L_p-norm, mutual proximity, distance measure, high dimensional dataset

- **Distance measures enhancements:** We highlight how common data science methods, such as the application of the mutual proximity, enhance the performance of distance measures.
- **Differentiation of epigenomic datasets using enhanced distance measurements:** We evaluate different distance measurements, both in their enhanced and non enhanced variants, in their performance to separate data obtained under the same experimental conditions (replicates), from data obtained under different experimental conditions.
- **Silhouette Index and Pearson correlation between distance matrix elements:** We outline why these indices are efficient to evaluate the performance of different distance measures in terms of cohesion and separability.

Introduction

In the last decade, advances in Next Generation Sequencing (NGS) technologies enabled the generation of a large number of epigenome tracks such as DNA methylation, chromatin modifications, transcription factor binding sites, DNA accessibility or transcription [1]. In most studies, multiple tracks are compared in order to obtain a better understanding on the interpretation of a genome sequence by the cell in specific contexts. Such comparisons commonly distinguish healthy cells from cancer cells or different cell types, occasionally at different differentiation stages [2–4]. These comparisons inevitably involve the choice of a distance or dissimilarity measure $d : V \times V \to \mathbb{R}^+$, a function that yields a non-negative real scalar, the distance between two high dimensional vectors containing the values of the genomic track at a given resolution.

Comparisons between performances of various distance measures, in terms of clusterability, have been carried out on

microarray data [5, 6]. Later on, such comparisons have been extended to next-generation sequencing (NGS) data, focusing today on single cell RNA-seq data. Kim et al. [7] studied the impact of four different distance measures on the clustering results of single cell RNA-Seq datasets and reported that the most efficient metric in such a case shall be Pearson's correlation. A following analysis by Skinnider et al. [8] found that Pearson's correlation could be outperformed by using two measures of proportionality described in [9].

Intrigued by papers published almost a decade ago that tackle the problem of finding optimal distance measures for high dimensional datasets [10–12] we systematically evaluated different distance measures and tried known methods such as the non-iterative contextual dissimilarity measure (NICDM) [10] or the mutual proximity (MP) [11] to further improve the distance measure.

In order to evaluate the performance of distance measures under various techniques that shall improve them, we take advantage of datasets that include multiple replicates. A replicate being data recorded under the same experimental conditions. To assess the quality of different distance measures we rely on the fact that experimental replicates are supposed to be almost identical and reasoned that a well-suited distance measure would minimize the distance between such replicates while maximizing the distance between the profiles determined under different experimental conditions. We outline that we designed our experiments relying only on replicate and non-replicate datasets avoiding the bias or distance measure dependency that clustering algorithms might introduce.

In cases where neither the MP nor the NICDM was applied to enhance the distance measures contrast our results are in agreement with the results obtained for single cell RNA-seq profiles [7] as we found that correlation-based distance measures (e.g. Pearson's correlation) outperformed distance-based metrics (e.g. Euclidean distance) for count profiles. Applying the MP generally enhances contrast and performance [11]. Further the MP raises the effectiveness to discern ensembles of replicate results for the L_1 norm to same level as correlation based distances under MP for mRNA-seq and miRNA-seq count profiles. This should yield an advantage in computational efficiency as L_1 can be implemented using fewer machine instructions than correlation based distance measures. Concerning binary profiles, that are for instance obtained after the application of a peak caller, we find that both, Yule and Jaccard distances, are performing well. The two binary distance measures again see an improvement as MP is applied.

Methods

Data collection

We obtained the epigenetic tracks annotating the human genome (hg38) from the data generated by the Canadian Epigenetics, Epigenomics, Environment and Health Research Consortium (CEEHRC)[1]. We downloaded the following datasets: Bisulfite-seq methylation data, H3K27ac, H3K27me3, H3K36me3, H3K4me3 and H3K9me3 histone ChIP-seq data, mRNA-seq, and miRNA-seq transcriptome data as well as plain input DNA control data. All tracks were obtained for different cell types, namely: CD19+, Basal, Glioma, Colorectal-normal, Thyroid-normal cells as well as healthy T-Cells. A overview of the obtained benchmark data is outlined in table 1. The data includes binary signals for the ChIP-seq experiments. These binary signals were generated by the CEEHRC consortium using a peak caller which is part of their data processing pipeline. For the Bisulfite-seq experiments we constructed two binary signals by introducing a manual cutoff of the provided fractional signal. The fractional signal is composed of values between 0 and 1. As we chose thresholds of 25% and 75% values below 0.25/0.75 were set to 0 and values above 0.25/0.75 to 1 respectively.

Data processing

All data has been extracted and ensembles of both 200 and 500 succeeding bases pairs were averaged. The means over 500 bases pairs allow us to verify the stability of our results. All further results in the main article are presented for tracks that were averaged to bins of 200 bases pairs. The relevant data for means over 500 bases pairs is shown in tables S3 and S4 in the supporting information, from where it it is clear that these results are equivalent to tables S1 and S2. Tables S1 and S2, also found in the supporting information, outline in detail results obtained for means over 200 bases pairs. libBigWig [13] was used for the extraction, averaging and binning purpose. Further data processing was performed using a set of custom in house written tools that are available to the pubic as C code[2]. These tools consist of sparse data handling routines, that allow us to avoid storing parts of the epigenomic dataset where no signal was recorded or where the signal equals to zero, saving valuable computer memory and speeding up the following implementations. All distance measures were carefully reimplemented by hand in their enhanced and non enhanced versions. Further the quality measurement indices outlined herein have been coded in C as well.

Distance measures

In the presented study we assessed the following distance measures for count profiles:

- L_p-norm: as defined by

$$d(x,y) = \left(\sum_i |x_i - y_i|^p\right)^{1/p}, \tag{1}$$

  implemented for $p = 1$, the Manhattan distance, $p = 2$, the Euclidean distance, $p = 3$, $p = 4$. Taking note that fractional norms in theory should perform better for high dimensional data we further used a value of $p = 0.5$ [14].
- Cosine distance:

$$d(x,y) = 1 - \sum_i \frac{x_i y_i}{\sqrt{x_i^2}\sqrt{y_i^2}}. \tag{2}$$

- Pearson distance: which is closely related to the Cosine distance and defined by:

$$d(x,y) = 1 - \sum_i \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(x_i - \bar{x})^2}\sqrt{(y_i - \bar{y})^2}}, \tag{3}$$

  where $\bar{x}$ and $\bar{y}$ are the means of the components in the vectors x and y.

[1] Information about CEEHRC and the participating investigators and institutions can be found at `http://www.cihr-irsc.gc.ca/e/43734.html`

[2] In house developed data processing code is published at `https://gitlab.in2p3.fr/mnhn-tools/distanceboost-mp-nc`

Table 1. The number of epigenetic tracks for each experiment and each cell type.

	Bisulfite[1]	mRNA[2,3]	miRNA[3]	H3K27ac	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3	Control[3]
CD19+	13	11	14	13	13	13	13	13	13	13
T-Cells	5	5	-	5	5	5	5	5	5	5
Basal	6	6	1	9	8	8	9	8	8	9
Glioma	12	11	12	12	12	12	12	12	12	12
Colorectal	15	15	15	15	15	15	15	15	15	15
Thyroid	9	7	9	9	9	9	9	9	9	9

[1]For binarisation prior to the application of Jaccard and Yule distance metrics data below either 25% and 75% methylation was set to zero and to one above.

[2]Only expression data read in direction from 3' to 5' was considered.

[3]No binarisation or peak calling was performed and hence no data is available for Yule and Jaccard distance measures.

Distances between binary signals were evaluated using the Jaccard and Yule measures:

- Jaccard distance:

$$d(x,y) = \frac{|a \cup b| - |a \cap b|}{|a \cup b|}, \tag{4}$$

- Yule distance:

$$\begin{aligned} a &= |\{i : i \in x, i \notin y\}|, \\ b &= |\{i : i \notin x, i \in y\}|, \\ c &= |\{i : i \in x, y\}|, \\ d &= |\{i : i \notin x, y\}|, \\ d(x,y) &= \frac{2ab}{cd}. \end{aligned} \tag{5}$$

In order to avoid floating point precision issues that may arise when summing a large number of elements, the Kahan summation correction was applied [15].

Enhanced Distance Measures

The MP introduced by [11], transforms a classic distance measure into a probability based distance measure improving the contrast at the same time. This method has the advantage that it does not require any prior knowledge of replicated experiments or data that forms close ensembles in the dataset. Knowing all pairwise distances between the different experiments in a dataset the MP is given by:

$$\mathrm{mp}(d(x,y)) = \frac{|\{i : d(x,i) > d(x,y)\} \cap \{i : d(y,i) > d(x,y)\}|}{N}, \tag{6}$$

which can be evaluated straightforward by counting the number of elements i having a distance to x and y greater than the distance $d(x,y)$. The mutual proximity $\mathrm{mp}(d(x,y))$ is itself a distance measure that increases the contrast in a dataset [11]. As the MP is probability based and normalized it can be used to linearly combine distance measures with different properties into a single distance measure:

$$d_{\mathrm{combined}} = \sum_j \alpha_j \mathrm{mp}(d_j(x,y)), \tag{7}$$

under the condition for the weighting factors to be $\sum_j \alpha_j = 1$. [11]. We used this property of the MP to create a mixed distance measure:

$$d(x,y) = \frac{1}{2}[\mathrm{mp}(d_J(x,y)) + \mathrm{mp}(d_Y(x,y))] \tag{8}$$

where $d_J(x,y)$ represents the Jaccard distance and $d_Y(x,y)$ the Yule distance.

The second method to enhance distance measures that we implemented is the Non-Iterative Contextual Dissimilarity Measure (NICDM) as first proposed in [10]:

$$\mathrm{NICDM}(d(x,y)) = d(x,y)\sqrt{\frac{\bar{r}^2}{r_x r_y}}, \tag{9}$$

where r_x is the mean of the distances in the ensemble, in our case epigenetic sequencing experiments that have been performed under the same experimental conditions, that x belongs to and $\bar{r}$ is the mean of all the means r_i for all ensembles. This method can either be used in an iterative manner, together with a clustering algorithm or requires prior knowledge of the dataset, in our case replicates of experiments. A property that might prove to be problematic in real world applications, where contrary to the data treated in this article, such a knowledge might not be available.

Our code implements the MP and NICDM as highlighted in equations (6) and (9), which we applied on all distance measures outlined herein.

Benchmark Indices

In order to estimate the quality of each distance measure we computed the *Silhouette* index [16] based on the close ensembles formed by the same cell types, defined here as experimental replicates. The Silhouette for a single pair of ensembles was implemented as follows:

$$\begin{aligned} k_{j,\mathrm{int}} &= \frac{1}{n_{\mathrm{int}}} \sum_i^{n_{\mathrm{int}}} d(c_{\mathrm{int}}(i), c_{\mathrm{int}}(j)) \\ k_{j,\mathrm{ext}} &= \frac{1}{n_{\mathrm{int}}} \sum_i^{n_{\mathrm{ext}}} d(c_{\mathrm{ext}}(i), c_{\mathrm{int}}(j)) \\ s_j &= \begin{cases} k_{j,\mathrm{int}} < k_{j,\mathrm{ext}} : & 1 - \frac{k_{j,\mathrm{int}}}{k_{j,\mathrm{ext}}} \\ k_{j,\mathrm{int}} = k_{j,\mathrm{ext}} : & 0 \\ k_{j,\mathrm{int}} > k_{j,\mathrm{ext}} : & \frac{k_{j,\mathrm{ext}}}{k_{j,\mathrm{int}}} - 1 \end{cases} \\ S &= \frac{1}{n_{\mathrm{int}}} \sum_j^{n_{\mathrm{int}}} s_j, \end{aligned} \tag{10}$$

where $c_{\mathrm{int/ext}}(i)$ is the i-th element (experiment under same experimental conditions) belonging to the *internal (same) /*

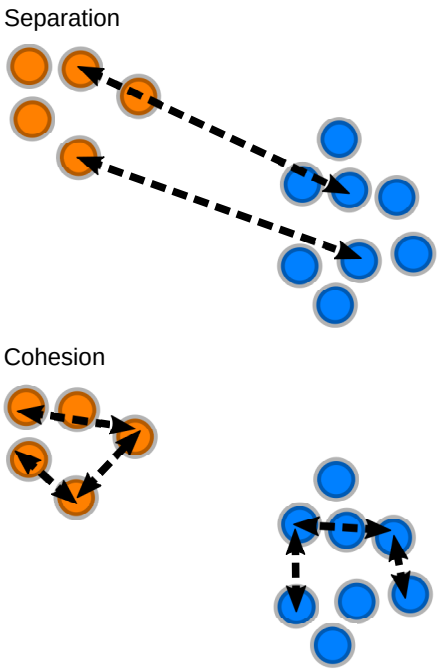

Fig. 1. Separation, measured by $k_{j,\text{ext}}$ and cohesion, measured by $k_{j,\text{int}}$, two wanted properties that the Silhouettes as outlined in equation (10) and the Silhouette Index in equation (11) describe. If a dataset contains an ensemble of datapoints, labeled to be of the same group that exerts cohesion within its members while it is separated from differently labeled datapoints, the Silhouette approaches $+1$.

external (different) cell type and n_{int} is the number of elements of the *internal* cell type. The Silhouette varies between $-1 \leq S \leq +1$ and is negative if the distances between elements belonging to the same cell type (internal elements) to those of a different cell type (external elements) are shorter then the distances between all the "internal" elements of the same cell type. If the Silhouette is positive the internal distances between the elements of the same cell type are smaller than the external distances to the elements of a different cell type. Therefore a high quality distance measure should yield a positive Silhouette index close to $+1$ [16]. We further make use of the Silhouette Index, the arithmetic mean of Silhouettes obtained for all cell type pairs (c_i, c_j) in a dataset:

$$SI = \frac{\sum_{\text{pairs}} S(c_i, c_j)}{n_{\text{pairs}}}, \tag{11}$$

Silhouettes, as well as the Silhouette Index are one of the methods of choice as they efficiently tackle the problem of comparing cohesion and separability. With cohesion representing the pairwise intra cell type distances, and separability the pairwise inter cell type distances as outlined in figure 1. A high quality distance measure is thus expected to minimize cohesion while it is at the same time maximizing separation.

To further improve our benchmarking capabilities we also compute the Pearson correlation P between the elements of a distance matrix $D_{i,j}(d(x_i, y_j))$ and the elements of a matrix $M_{i,j}$ where:

$$M_{i,j} = \begin{cases} 0: & (x_i \in c_i) \wedge (y_j \in c_i) \\ 1: & \text{otherwise} \end{cases}. \tag{12}$$

c_i in this equation corresponds to different ensembles that group together replicate experiments belonging to the same cell types, indexed by i. Thus the matrix is 0 if x and y are of the same cell type and 1 otherwise. The matrix $M_{i,j}$ is especially interesting as it represents the perfect MP for the perfect distance measure. The matrix $M_{i,j}$ represents a perfect distance matrix, built from a hypothesized perfect distance metric where all intra ensemble pairwise distances, distances between experiments of the same cell type, yield 0, while pairwise inter ensemble distances and hence, distances concerning experiments of two different cell types, yield 1. The Pearson correlation between the matrix elements of the distance matrix $D_{i,j}$, dependent on the distance metric used, and $M_{i,j}$:

$$P = \sum_{\forall i,j} \frac{M_{i,j} D_{i,j}(d(x_i, y_j))}{\sqrt{(M_{i,j} - \bar{M})^2}\sqrt{(D_{i,j} - \bar{D})^2}}, \tag{13}$$

with $\bar{M}$ and $\bar{D}$ being the arithmetic means of matrix elements, is thus our second benchmark index. P shall just as SI approach $+1$ the better suited the distance measure is. This P benchmark index is directly derived from commonly used Mantel statistics [17] which is frequently used to compare different distance measures. The sole differences are:

1. As we perfectly know our dataset, matrix elements $D(i,j)$ and $M(i,j)$ in both matrices provide distances between the same profiles i, j. Therefore we can completely avoid the critiqued (c.f. G. Guillot and F. Rousset [18]) randomization/permutation procedure that normally is applied on one of the matrices.
2. We compare the distance matrix obtained from a certain distance measure to a fabricated perfect distance matrix that has the properties of an optimal distance measure applied on our dataset $M(i,j)$. Traditional Mantel Statistics on the other hand compares empirically measured distance matrices [17] where knowledge about such an optimal distance matrix $M(i,j)$ is not available.

We point out that the SI and P indices can be computed without performing clustering and hence, allow us to obtain an unbiased view on the subject if the MP or the NICDM enhance distance measures used for epigenetic profiles or not. Further this approach allows us to find the best combinations of enhancement functions coupled with classical distance measures. The study was purposeful designed in this way in order to evaluate the effects of different distance measures independent from clustering algorithms.

Benchmark indices for the binary profiles of the Bisulfite-seq experiment are calculated by taking the mean $SI = \frac{1}{2}[SI(25\%)+SI(75\%)]$ and $P = \frac{1}{2}[P(25\%)+P(75\%)]$ for profiles generated from a 25% and 75% cutoff of the fractional data provided by the CEEHRC, CIHR consortiums.

Results

In order to evaluate distance measures for all the variety of experiments that can be found in the epigenomic toolbox, we used datasets for DNA methylation (Bisulfite-seq), gene expression (mRNA-seq), microRNA (miRNA-seq), localized histone modifications (H3K27Ac, H3K4me1, H3K4me3 ChIP-seq) and spreading histone modifications (H3K27me3, H3K9me3, H3K36me3 ChIP-seq) as demonstrated in table 1. For all these datasets, we evaluated the following dissimilarity measures: Count profiles were differentiated using

the L_p distance with $p = 0.5, 1, 2, 3, 4$, the Cosine and Pearson dissimilarity, respectively outlined in equations (1), (2), (3). Further binary profile distances were evaluated using the Jaccard and Yule distances, as described by equations (4) and (5). For each dissimilarity measure we also applied both the MP or the NICDM procedures that transforms the distance matrix for all datasets found in table 1. In order to assess the ability of each measure, with or without the application of the MP and NCDIM procedures, to discriminate profiles that come from the replicates of the same experiment with other profiles, we computed two indices: The Silhouette index SI and the Pearson index P which are outlined in equations (11) and (13). The mean results with and without the application of either the MP or NICDM are highlighted in table 2 and 3. Table 2 outlines the performance according to the P index: Here the application of the NICDM only provided an improvement on the L_1, L_2, L_3, L_4 as well as the Jaccard and Yule distance measures. Under the same conditions the MP provided improvements for all distance measures, outperforming the NICDM in most cases. In the same manner table 3 outlines how all distance measures see a performance increase according to the SI index if the MP or the NICDM is applied. We however do not see a clear advantage of using the NICDM in comparison to the more flexible MP. For in depth details we refer the reader to tables S1 and S2 in the supplemental material where exact values for each dataset are shown. From tables S1 and S2 we further assembled Figure 2 which highlights the results for all distance/dissimilarity measures with and without application of the MP. In these diagrams best performing distance measures find themselves in the upper right corner, while weak performing distance measurements are found around the origin in the lower left corner. A clear shift to the upper right corner is visible as the MP is applied onto the different distance measures. In the case of the Yule and Jaccard distances the effect on our benchmark indices that are bound to the interval [0,1] can yield a boost of 0.3, greatly increasing the effectiveness of the distance measure in terms of separation and cohesion.

Finally we would like to point out that our distance measures, in both MP improved and non-improved form, applied on binary data, either obtained through a simple threshold as in the case of bisulfite-seq, or as provided by the datasets processed with the peak calling pipeline of the CEEHRC consortium, performed better then simple count or intensity profiles. This could be explained by a feature selection during the binarisation process. The highest scores for the SI and P indices were obtained with either bisulfite-seq or H3K27ac CHiP-seq experiments, underlying the superiority of these assays to discriminate between cell types. mRNA-seq counts and normalized datasets were also found to perform well at this task in cases where the L_1, Pearson or Cosine dissimilarity measures enhanced by MP were used. This result further holds for miRNA experiments. Concerning histone modifications, we found that local modifications (H3K27Ac, H3K4me1 and H3K4me3) are better suited to discriminate profiles from different cell types than propagating tri-methyl marks (H3K36me3, H3K27me3 and H3K9me3). As expected, the control datasets for ChIP-seq experiments were not efficient in separating groups of experimental replicates.

Intriguingly the Yule distance performs well in terms of the SI index but is much less efficient in terms of our P index. For the Jaccard distance the effect seems to be reversed, and the distance performs better for P than for SI. This lead us to the idea to combine both distances using the MP as shown in equation (8). In both cases applying the MP on the two measures, Jaccard and Yule, without combining them however yielded almost the same results, improving the weak indices for both distance measures so that both distance measures where rejoining themselves exerting almost the same performance.

Discussion

The presented results highlight how classical distance measures that are used to discern individual epigenetic profiles can be to a great extend improved by applying the MP. In cases where peak finding or binarisation is possible the Yule and Jaccard distances are leading combined with the application of the MP to the best results. However certain datasets such as those from mRNA-seq experiments are not well suited for binarisation. In such cases, we refer the reader to figure 2 in order to find the best distance measure for the application. L_1, Pearson or Cosine, distances again together with MP seem to perform best on high dimensional datasets of this kind. We could not find an advantage using fractional norms as suggested in [14]. In several cases, especially when applied to binary data, the MP shows a drastic increase of performance and we strongly recommend its usage in order to increase contrast as one compares epigenomic signals. We did not find a clear advantage of the NICDM over MP, and see its application difficult as it requires prior knowledge that one might want to discover in the dataset in using the distance measure, i.e. clustering the dataset. The MP can further be used to combine the properties of different distance measures, which we tried for the Yule and the Jaccard distance. As both the Yule and the Jaccard distance by themselves underwent a strong performance increase as we applied the MP, a mixture of both as outlined by equation (8) did not yield better results.

Competing interests

There is NO Competing Interest.

Author contributions statement

T.H., J.B.M. and J.M. proposed the outlined study. T.H., J.B.M and L.C. performed the numerical experiments. T.H. and J.M. wrote the article manuscript.

Acknowledgments

This work has been funded by the French Muséum Nationale d'Histoire Naturelle (MNHN) and Sorbonne Université.

References

1. Wei Xie, Matthew D Schultz, Ryan Lister, Zhonggang Hou, Nisha Rajagopal, Pradipta Ray, John W Whitaker, Shulan Tian, R David Hawkins, Danny Leung, et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, 153(5):1134–1148, 2013.
2. Jason D Buenrostro, M Ryan Corces, Caleb A Lareau, Beijing Wu, Alicia N Schep, Martin J Aryee, Ravindra Majeti, Howard Y Chang, and William J Greenleaf. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell*, 2018.
3. M Ryan Corces, Jason D Buenrostro, Beijing Wu, Peyton G Greenside, Steven M Chan, Julie L Koenig, Michael P

Table 2. Performance according to the element wise Pearson correlation of the distance matrix to a perfect distance matrix in means over datasets ± the standard errors

	L2	L1	L3	L4	L(1/2)	Cosine	Pearson	Yule	Jaccard	Y+J
direct	0.27±0.01	0.37±0.04	0.20±0.01	0.17±0.01	0.37±0.03	0.45±0.06	0.45±0.06	0.57±0.05	0.35±0.04	
mp	0.41±0.03	0.46±0.04	0.33±0.02	0.30±0.02	0.41±0.04	0.52±0.05	0.52±0.05	0.63±0.04	0.65±0.04	0.65±0.04
nc	0.31±0.02	0.38±0.04	0.24±0.02	0.22±0.02	0.36±0.03	0.44±0.05	0.44±0.05	0.60±0.03	0.45±0.04	
Δmp	0.14±0.02	0.09±0.01	0.14±0.02	0.13±0.02	0.05±0.01	0.07±0.02	0.06±0.02	0.06±0.04	0.30±0.04	
Δnc	0.04±0.01	0.01±0.01	0.04±0.01	0.04±0.01	-0.01±0.00	-0.01±0.02	-0.01±0.02	0.03±0.04	0.10±0.04	

mp Distance measures that MP has been applied on.

nc Distance measures that NCIDM has been applied on.

Δ Difference between the distance measure with and without MP or NICDM.

Table 3. Performance according to the Silhouette index in means over datasets ± the standard errors.

	L2	L1	L3	L4	L(1/2)	Cosine	Pearson	Yule	Jaccard	Y+J
direct	0.32±0.04	0.31±0.04	0.29±0.04	0.28±0.04	0.22±0.03	0.42±0.05	0.42±0.05	0.34±0.04	0.60±0.07	
mp	0.41±0.05	0.45±0.06	0.35±0.04	0.32±0.04	0.40±0.05	0.50±0.06	0.50±0.06	0.61±0.06	0.62±0.06	0.62±0.06
nc	0.35±0.03	0.34±0.04	0.31±0.03	0.30±0.03	0.24±0.03	0.50±0.04	0.50±0.04	0.37±0.03	0.69±0.04	
Δmp	0.09±0.02	0.12±0.02	0.06±0.02	0.05±0.02	0.18±0.03	0.07±0.02	0.07±0.02	0.27±0.03	0.02±0.03	
Δnc	0.03±0.02	0.04±0.02	0.03±0.02	0.03±0.03	0.02±0.01	0.07±0.02	0.07±0.02	0.02±0.02	0.09±0.04	

mp Distance measures that MP has been applied on.

nc Distance measures that NCIDM has been applied on.

Δ Difference between the distance measure with and without MP or NICDM.

Snyder, Jonathan K Pritchard, Anshul Kundaje, William J Greenleaf, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature genetics*, 48(10):1193, 2016.

4. Andrew B Stergachis, Shane Neph, Alex Reynolds, Richard Humbert, Brady Miller, Sharon L Paige, Benjamin Vernot, Jeffrey B Cheng, Robert E Thurman, Richard Sandstrom, et al. Developmental fate and cellular maturity encoded in human regulatory dna landscapes. *Cell*, 154(4):888–903, 2013.
5. R. Gentleman, B. Ding, S. Dudoit, and J. Ibrahim. *Distance Measures in DNA Microarray Data Analysis*, page 189–208. Springer New York, New York, NY, 2005.
6. Raffaele Giancarlo, Giosuè Lo Bosco, and Luca Pinello. Distance functions, clustering algorithms and microarray data analysis. In *Proceedings of the 4th International Conference on Learning and Intelligent Optimization*, LION'10, pages 125–138, Berlin, Heidelberg, 2010. Springer-Verlag.
7. Taiyun Kim, Irene Rui Chen, Yingxin Lin, Andy Yi-Yang Wang, Jean Yee Hwa Yang, and Pengyi Yang. Impact of similarity metrics on single-cell rna-seq data clustering. *Briefings in bioinformatics*, 2018.
8. Skinnider Michael A., Squair Jordan W., and Foster Leonard J. Evaluating measures of association for single-cell transcriptomics. *Nature Methods*, 16(5):381–386, 2019.
9. Quinn Thomas P., Richardson Mark F., Lovell David, and Crowley Tamsyn M. propr: An r-package for identifying proportionally abundant features using compositional data analysis. *Scientific Reports*, 7(1):16252, 2017.
10. H. Jegou, H. Harzallah, and C. Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.
11. Dominik Schnitzer, Arthur Flexer, Markus Schedl, and Gerhard Widmer. Local and global scaling reduce hubs in space. *Journal of Machine Learning Research*, 13(10), 2012.
12. Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering, 2004. *Advances in Neural Information Processing Systems*, 17, 2005.
13. Fidel Ramírez, Devon P Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44(W1):W160–W165, 04 2016.
14. Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional spaces. In *Proceedings of the 8th International Conference on Database Theory*, ICDT '01, page 420–434, Berlin, Heidelberg, 2001. Springer-Verlag.
15. W. Kahan. Pracniques: Further remarks on reducing truncation errors. *Commun. ACM*, 8(1):40, January 1965.
16. Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.
17. Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2 Part 1):209–220, 1967.
18. Gilles Guillot and François Rousset. Dismantling the mantel tests. *Methods in Ecology and Evolution*, 4(4):336–344, 2013.

Thomas Haschka teaches Machine Learning (ML) and Artificial Intelligence (AI) at the French Natural History

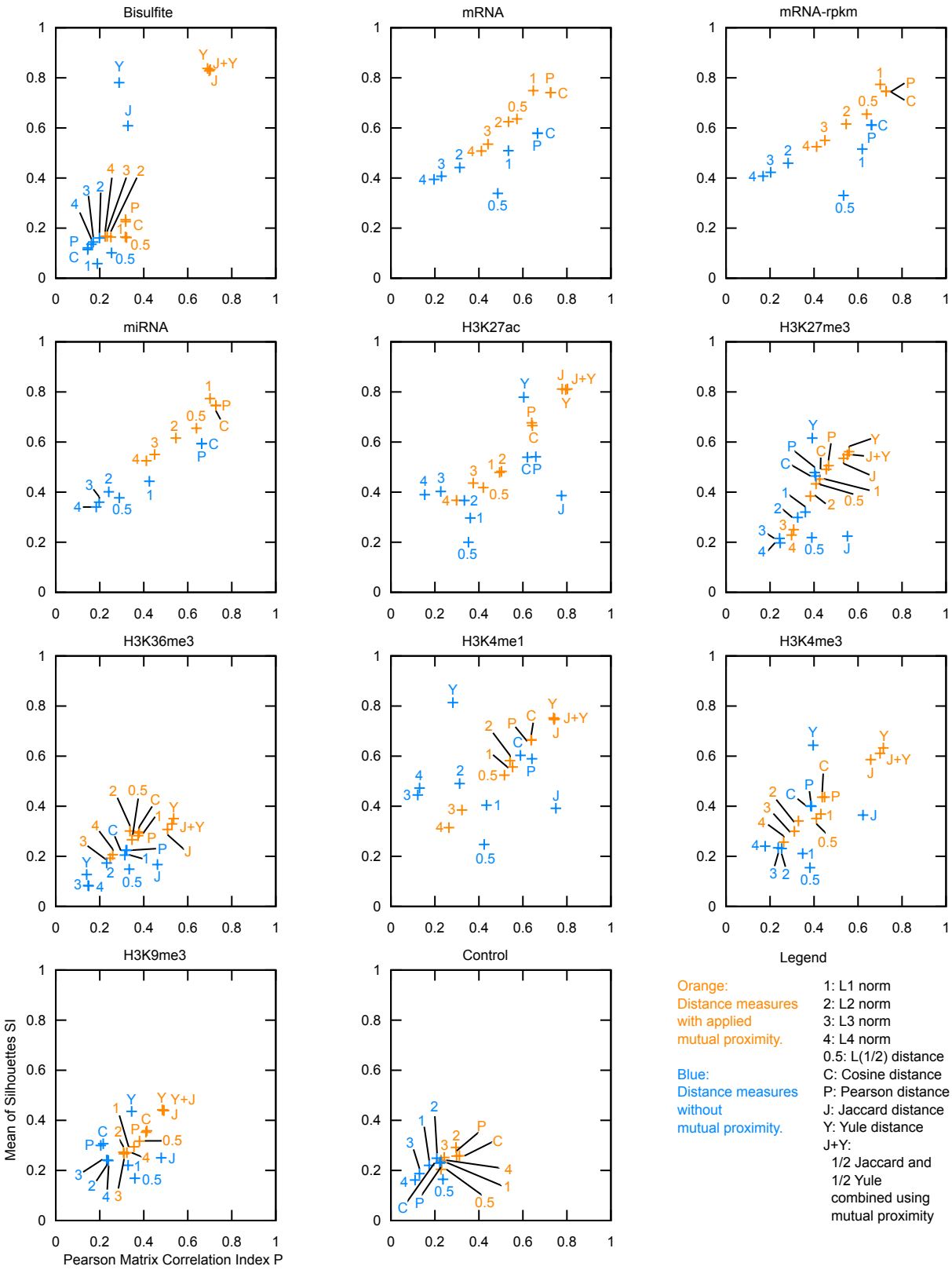

Fig. 2. The performance of different distance measures under the Silhouette Index SI and the Pearson correlation of distance matrix elements to a perfect distance matrix P as shown in equations (11) and (13) respectively.

Museum (MNHN) in Paris. In 2012 he defended his PhD theses in molecular biophysics under the supervision of Manuel Dauchez at the University of Reims Champagne Ardenne. His research topics currently revolve around the application of AI to molecular evolution and phylogenetics.

Jean-Baptiste Morlot is the Chief Technical Officer of DeepLife in Paris. He was trained as a theoretical physicist and received his PhD in machine learning applied to sequencing data. He has been working since then on multi-omics cell modelling by integrating single cell and bulk sequencing data. His main current interest is to develop computational models that integrate diverse data sources to better understand and treat diseases.

Léopold Carron is a Post doctoral research scientist at the Laboratory of Computational and Quantitative Biology at the Sorbonne in Paris. He made his PhD under the supervision of Julien Mozziconacci in bioinformatics. He works on DNA space organization, more specially in developing new tools to understand the link between epigenomics and 3D conformation folding.

Julien Mozziconacci

is a professor at the National Museum of Natural History in Paris. He was trained as a theoretical physicist and received his PhD in Physics on the mutli-scale architecture of chromosomes. He has been working since then on the modeling of chromosomal structure, from nucleosomes to whole nuclei, by integrating various experimental approaches. Currently his main interests are the role of DNA repeats in the functioning of the genome as well as in evolution. Further he is interested in the application of deep learning to DNA sequences.