Similarity and dissimilarity metrics

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

From profiles to (dis)similarity matrix

Expression profiles (gene/sample)

	Sample 1	Sample2	...	Sample j	...	Sample p
Gene 1	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}
Gene 2	x_{21}	x_{22}	...	x_{2j}	...	x_{2p}
...	...	...	...	...	...	...
Gene i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ip}
...	...	...	...	...	...	...
Gene n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{np}

(Dis)similarity matrix (gene/gene)

	Gene 1	Gene 2	...	Gene j	...	Gene n
Gene 1	d_{11}	d_{12}	...	d_{1j}	...	d_{1p}
Gene 2	d_{21}	d_{22}	...	d_{2j}	...	d_{2p}
...	...	...	...	...	...	...
Gene i	d_{i1}	d_{i2}	...	d_{ij}	...	d_{ip}
...	...	...	...	...	...	...
Gene n	d_{n1}	d_{n2}	...	d_{nj}	...	d_{np}

(Dis)similarity matrix (sample/sample)

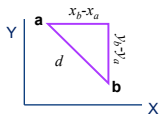
	Sample 1	Sample 2	...	Sample j	...	Sample n
Sample 1	d_{11}	d_{12}	...	d_{1j}	...	d_{1p}
Sample 2	d_{21}	d_{22}	...	d_{2j}	...	d_{2p}
...	...	...	...	...	...	...
Sample i	d_{i1}	d_{i2}	...	d_{ij}	...	d_{ip}
...	...	...	...	...	...	...
Sample n	d_{n1}	d_{n2}	...	d_{nj}	...	d_{np}

Choice of a (dis)similarity metric

- A crucial parameter for classification is the choice of an appropriate metrics to measure the similarity or dissimilarity between objects
- There are plenty of (dis)similarity metrics
- To cite a few:
  - Euclidian distance
  - Manhattan distance
  - Pearson's coefficient of correlation
  - Mahalanobis distance
  - $\chi 2$ distance
- The choice of the metrics depends on the data type

Euclidian distance

- We are familiar with the Euclidian distance in a 2-dimensional space.



$$D_{ab} = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$

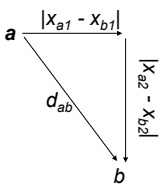
- The concept naturally extends to spaces with higher dimension (p-dimensional space)
  - X_{ai} and X_{bi} are the values of the i^{th} dimension (variable) to the objects a and b, respectively.
  - p is the number of dimensions

$$D_{ab} = \sqrt{\sum_{j=1}^{p} (x_{aj} - x_{bj})^2}$$

- Two typical applications
  - The distance between genes is calculated in the space of samples (microarray chips)
  - The distance between samples is calculated in the space of genes (probes)

Weighted Euclidian distance

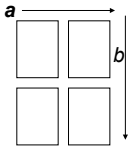
- The weighted Euclidian distance between two points is calculated as the Euclidian distance, with a specific weight w_j associated to each dimension j
  - a,b two points in the multi-variate space
  - p number of dimensions
  - w_j weight if the j^{th} dimension



$$D_{ab} = \sqrt{\sum_{j=1}^{p} w_i (x_{aj} - x_{bj})^2}$$

Manhattan distance

- The Manhattan distance between two points a and b is the weighted sum of the absolute differences in each dimension.
  - a,b two points in the multi-variate space
  - p number of dimensions
  - w_i weight if the i^{th} dimension



$$D_{ab} = \sum_{j=1}^{p} w_j |x_{aj} - x_{bj}|$$

Minkowski metrics

- The Minkowski metrics are a family of dissimilarity metrics, which can be tuned by a parameter (λ).
- Some particular values of λ give the metrics seen before.
  - λ=1 Manhattan distance
  - λ=2 Euclidian distance

$$D_{ab} = \sqrt[\lambda]{\sum_{j=1}^{p} w_j^{\lambda} \left| x_{aj} - x_{bj} \right|^{\lambda}}$$

Correlation-related metrics

- A detailed description of those metrics has been given in the chapter "**Correlation analysis**".
- The coefficient of correlation and several related metrics can be converted to dissimilarity metrics.
  - *mdp* mean dot product
  - *cor* correlation
  - *Ucor* uncentered correlation

$$mdp_{ab} = \frac{1}{p} \mathbf{x_a} \cdot \mathbf{x_b} = \frac{1}{p} \sum_{i=1}^{p} \left(x_{ai} \cdot x_{bi} \right)$$

$$mdpd_{ab} = k - dmp_{ab}$$

$$cor_{ab} = \frac{1}{p} \sum_{i=1}^{p} \left(\frac{x_{ai} - \hat{m}_a}{\hat{\sigma}_a} \right) \left(\frac{x_{bi} - \hat{m}_b}{\hat{\sigma}_b} \right)$$

$$Dcor_{ab} = 1 - cor_{ab}$$

$$Ucor_{ab} = \frac{\sum_{i=1}^{p} \left(x_{ai} . x_{bi} \right)}{\sqrt{\sum_{j=1}^{p} x_{aj}^{\ 2}} \sqrt{\sum_{j=1}^{p} x_{bj}^{\ 2}}}$$

Pearson's coefficient of correlation

- Pearson's coefficient of correlation is a classical metric of similarity between two objects.

$$c_P = \sum_{j=1}^{p} \left(\frac{x_{aj} - m_a}{\sigma_a} \right) \left(\frac{x_{bj} - m_b}{\sigma_b} \right) = \sum_{j=1}^{p} z_a z_b$$

- Where
  - a is the index of an object (e.g. a gene)
  - b is the index of another object (e.g. a gene)
  - i is an index of dimension (e.g. a chip)
  - m_i is the mean value of the i^{th} dimension
- Note the correspondence with z-scores.
- The correlation is comprised between -1 and 1.
- It can be converted to a **correlation distance** by a simple operation.

$$d_P = 1 - c_P$$

Spearman's rank correlation coefficient

- Spearan's correlation is computed by
  - calculating the rank of each object along each variable and
  - computing Pearson's correlation between the rank values.
- Properties
  - Robust to the presence of outliers (values that strongly differ from the rest of the population).
    - This property may be interesting for microarray measurements, where the presence of noise can lead to outliers.
  - Insensitive to the linearity of the relationship between variables.
  - Particlar case: if two variables are monotonically related , Spearman's coefficient is 1.

Generalized coefficient of correlation

- This metrics was proposed by Mike Eisen in the first article describing a clustering method applied to gene expression profiles (Eisen et al., 1998).
- Pearson correlation can be generalized by using an arbitrary reference (*r*).
- Pearson's correlation is a particular case where $r_a=m_a$ and $r_b=m_b$.

$$c_P = \sum_{j=1}^{p} \left(\frac{x_{aj} - r_a}{\sqrt{\frac{1}{p} \sum_{k=1}^{p} \left(x_{ak} - r_a \right)^2}} \right) \left(\frac{x_{bj} - r_b}{\sqrt{\frac{1}{p} \sum_{k=1}^{p} \left(x_{bk} - r_b \right)^2}} \right)$$

Uncentred correlation

- Another particular case of the generalized correlation: one can arbitrarily take the value *r=0* as reference. This is called the uncentred correlation.
- This choice can be relevant if the object is a gene, and the value 0 represents non-regulation.

$$c_P = \sum_{j=1}^{p} \left(\frac{x_{aj}}{\sqrt{\frac{1}{p} \sum_{k=1}^{p} x_{ak}^{\ 2}}} \right) \left(\frac{x_{bj}}{\sqrt{\frac{1}{p} \sum_{k=1}^{p} x_{bk}^{\ 2}}} \right)$$

Dot product

- In principle, the dot product combines advantages of the Euclidian distance and of the coefficient of correlation
  - It takes positive values to represent co-regulation, and negative values to represent anti-regulation (as the coefficient of correlation)
  - It reflects the strength of the regulation of both genes, since it uses the real values (as the Euclidian distance) rather than the standardized ones (as the coefficient of correlation).
- It is not because the dot product seems a good metric in principle that it will be good in practice. This has to be evaluated on the basis of some testing set, for which the classes are known.

$$dp = \sum_{j=1}^{p} \left(x_{aj} * x_{bj} \right)$$

Impact of the choice of similarity and dissimilarity metrics

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

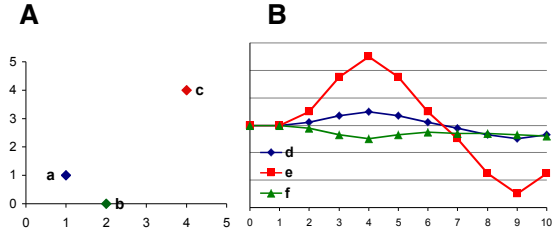
Choice of a metric for the clustering of gene expression data

- Euclidian distance
  - takes into account the absolute level of regulation (provided the genes have not been standardized)
  - Does not distinguish anti-correlation from absence of correlation
- Pearson's correlation
  - Indicates anti-correlation as well as correlation.
  - Does not indicate the absolute level of regulation.
  - Problem : assumes that the reference for each gene is the mean of its profile -> implicitly, one consider that each gene is on the average not regulated in the data set.
- Uncentered correlation
  - Indicates anti-correlation as well as correlation.
  - Does not indicate the absolute level of regulation.
  - Assumes that the reference level is 0, i.e. the level of the control experiment (the contribution of the green measurement to the log ratio)
- Dot product
  - In principle, the dot product combines advantages of the Euclidian distance and of the coefficient of correlation
  - It takes positive values to represent co-regulation, and negative values to represent anti-regulation (as the coefficient of correlation)
  - It reflects the strength of the regulation of both genes, since it uses the real values (as the Euclidian distance) rather than the standardized ones (as the coefficient of correlation).

Choice of a metric for the clustering of gene expression data

- Warning: it is not because the dot product seems a good metric in principle that it will be good in practice.
- This has to be evaluated on the basis of some testing set, for which the classes are known. This evaluation must be done on a case-per-case basis.
- It can easily be conceived that a metric could be appropriate for the clustering of columns (samples) and another metric for the clustering of rows (genes)
  - The coefficient of correlation seems a reasonable metric to compare two samples.
  - To compare two genes, it makes a strong (and generally not valid) assumption: the centering around the mean implicitly means that one considers that each gene is on the average not regulated in the set of experiments.

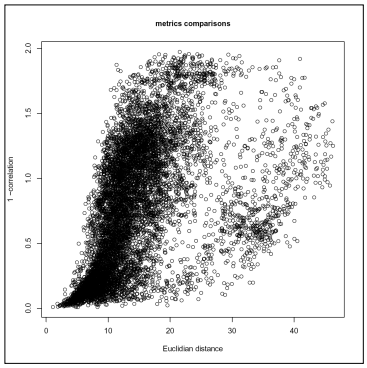
Impact of the distance metrics



Euclidian distances
• **a** close to **b**
Correlation coefficient
• **a** close to **c**

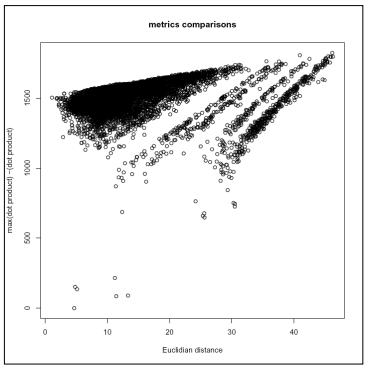
Euclidian distances
• **d** closer to **f** than to **e**
Correlation coefficient
• **d** closer to **e** than to **f**

Metrics comparison - carbon sources



- On this figure, each dot represents a pair of genes from the carbon source experiment.
- We selected the 133 genes showing a significant response in at least one of the 13 chips.
- For each pair of genes, we calculated the **Euclidian distance** (X axis) and Pearson's **centred coefficient of correlation** (Y axis).
- The plot shows that the two metrics are related but distinct.
- The cloud of points seems to be inhomogeneous: there are at least two separate trends.

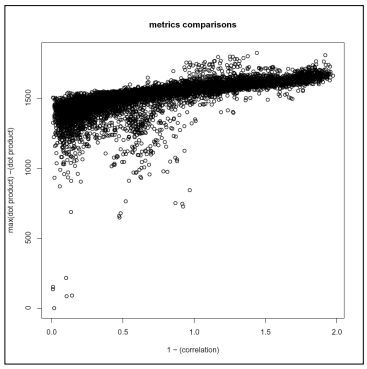
Metrics comparison - carbon sources



- **Euclidian distance** and **dot product** are related but there are be several apparent groups of gene pairs.
- Why ???????

Metrics comparison - carbon sources

- **Coefficient of correlation (centred)** and **dot product.**



Statistics Applied to Bioinformatics

Poisson-based similarity and dissimilarity metrics

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

Introduction

- The classical metrics of (dis)similarity can be used in a large number of contexts, but in some case they are not optimal.
- For example, they are not appropriate for calculating distances between discrete variables, such as motif occurrences in cis-regulatory sequences.
- In 2004, we proposed a series of metrics to address this specific purpose, and compared their performance with that of the classical metrics.

- **Note (2010):** with the advent of Next Generation Sequencing (NGS), transcriptome is measured by counting the number of reads sequenced for each mRNA of various samples. In the near future, it might be useful to come back to those Poisson-based metrics and to develop other statistics dedicated to the analysis of count-based data.

- van Helden, J. (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics, 20, 399-406.

Poisson-based similarity metric

- The probability to observe at least *x* common occurrence of pattern *i* in sequences *a* and *b* is the joint probability of observing at least *x* occurrences in sequence *a* and at least *x* occurrences in sequence *b*.

$$\begin{array}{ll} C_i^{ab} > 0 & P(x \geq C_i^{ab}) = \left[1 - F\left(C_i^{ab} - 1, m_i\right)\right]^2 \\ C_i^{ab} = 0 & P(x \geq C_i^{ab}) = 1 \end{array}$$

- Lower probabilities correspond to higher similarities. The probability of common occurrences can be converted in a similarity metrics.

$$s_i^{ab} = 1 - P(x \geq C_i^{ab})$$

- van Helden, J. (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics, 20, 399-406.

Multi-variate Poisson-based similarity

- A multi-variate similarity metric can be calculated as the average of single-variate metrics :

$$S_{add}^{ab} = \frac{1}{p} \sum_{i=1}^{p} s_i^{ab}$$

- Alternatively, one can consider the geometric mean, which reflects the joint probability of common occurrences for the different patterns :

$$S_{prod}^{ab} = 1 - \sqrt[p]{\prod_{i=1}^{p} P(x \geq C_i^{ab})}$$

- van Helden, J. (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics, 20, 399-406.

Poisson-based dissimilarity

$$d^{ab}_{distinct_i} = \left| F(N^b_i, m_i) - F(N^a_i, m_i) \right|$$

$$D^{ab}_{distinct} = \frac{1}{p} \sum_{i=1}^{p} d^{ab}_i$$

- van Helden, J. (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics, 20, 399-406.

Poisson-based dissimilarity based on over-representation

$$d^{ab}_{over_i} = \left| P(x \geq N^a_i) - P(x \geq N^b_i) \right| = \left| F(N^b_i - 1, m_i) - F(N^a_i - 1, m_i) \right|$$

$$D^{ab}_{over} = \frac{1}{p} \sum_{i=1}^{p} d^{ab}_i$$

- van Helden, J. (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics, 20, 399-406.