On the use of self-organizing maps for the representation of Barcoding data : an application to *Hylomyscus* data

M. Olteanu[1], V. Nicolas[2], C. Laredo[3]

[1] Laboratoire SAMM, EA 4543, Université Paris 1, France
[2] UMR 5202, Muséum National d'Histoire Naturelle, Paris, France
[3] Laboratoire MIA, Jouy en Josas, INRA, France

Methods

➢ Why use projection algorithms ?

o Visualizing high dimensional data in two dimensions
o Insight on the proximities between samples
o Linear projection with PCA
o Non-linear projection algorithms (SOM, MDS, ISOMAP,...)

Self-organizing maps (SOM)

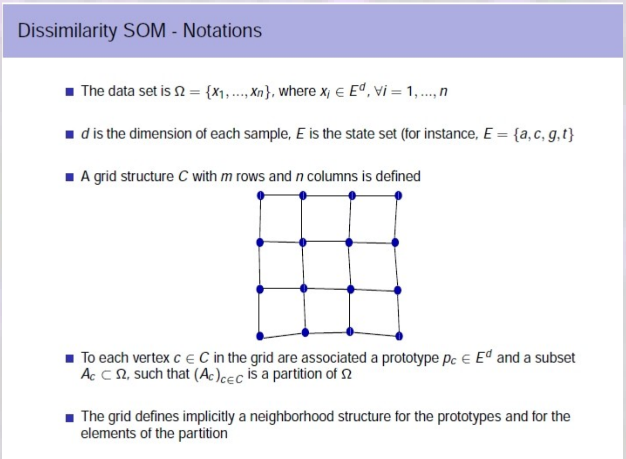
o Initially designed for Euclidean data (Koh2001)
o Neighborhood structure, topology preservation
o Vector quantization (clustering, unsupervised classification)
o Stochastic algorithm
o Generalization of k-means

SOM for dissimilarity data (Con2006)

o Very general method
o Euclidean distance is replaced by a dissimilarity measure

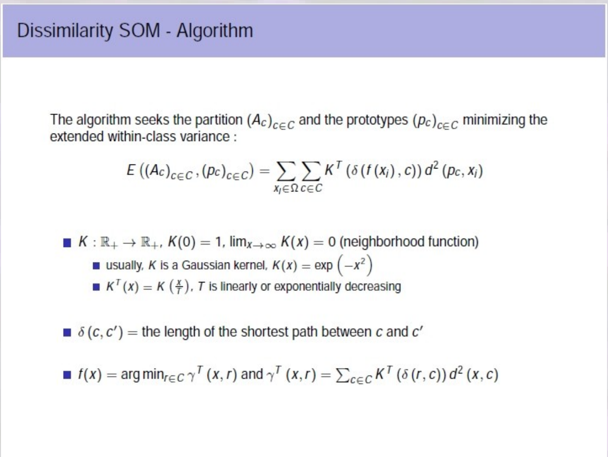
No hypothesis on the structure of the data

➢ « No free lunch »
o Batch algorithm
o Important computation time
o Sensitive to the intialization



Dissimilarity SOM - Notations

- The data set is $\Omega = \{x_1, ..., x_n\}$, where $x_i \in E^d$, $\forall i = 1, ..., n$
- d is the dimension of each sample, E is the state set (for instance, $E = \{a, c, g, t\}$
- A grid structure C with m rows and n columns is defined
- To each vertex $c \in C$ in the grid are associated a prototype $p_c \in E^d$ and a subset $A_c \subset \Omega$, such that $(A_c)_{c \in C}$ is a partition of Ω
- The grid defines implicitly a neighborhood structure for the prototypes and for the elements of the partition



Dissimilarity SOM - Algorithm

The algorithm seeks the partition $(A_c)_{c \in C}$ and the prototypes $(p_c)_{c \in C}$ minimizing the extended within-class variance :

$$E\left((A_c)_{c \in C}, (p_c)_{c \in C}\right) = \sum_{x_i \in \Omega} \sum_{c \in C} K^T\left(\delta\left(f\left(x_i\right), c\right)\right) d^2\left(p_c, x_i\right)$$

- $K : \mathbb{R}_+ \to \mathbb{R}_+$, $K(0) = 1$, $\lim_{x \to \infty} K(x) = 0$ (neighborhood function)
  - usually, K is a Gaussian kernel, $K(x) = \exp\left(-x^2\right)$
  - $K^T(x) = K\left(\frac{x}{T}\right)$, T is linearly or exponentially decreasing
- $\delta(c, c') =$ the length of the shortest path between c and c'
- $f(x) = \arg\min_{r \in C} \gamma^T(x, r)$ and $\gamma^T(x, r) = \sum_{c \in C} K^T\left(\delta(r, c)\right) d^2(x, c)$

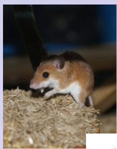
References

1. T. Kohonen (2001) Self-organizing maps, Springer

2. B. Conan-Guez, F. Rossi, A. El Golli (2006) Fast algorithm and implementation of dissimilarity self-organizing maps, Neural Networks, 19(6-7), p. 855-863

Data set

➢ Hylomyscus tribe (Murinae family)

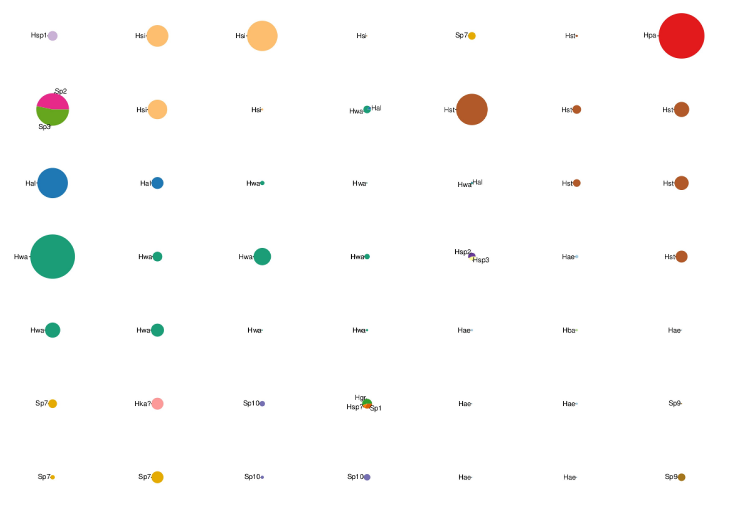


o 482 samples
o Gene Cytb (1267 sites)
o Non-polymorphic sites and sites with more than 20% missing data were deleted (421 sites)
o Some samples are well identified and characterized
o Some samples are not labeled with any species

➢ Comparison of two samples

o Kimura-2P dissimilarity

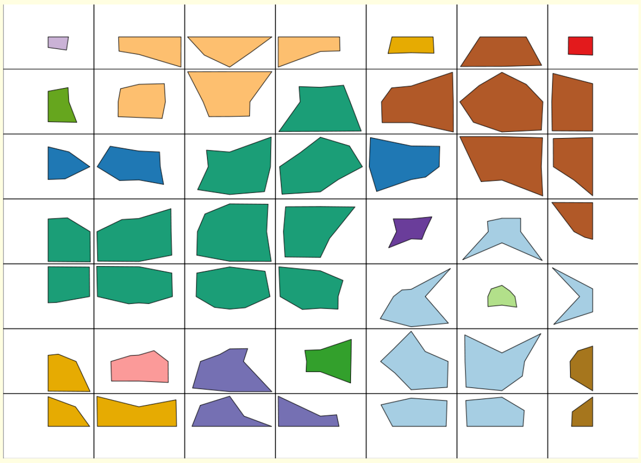
A posteriori highlighting the available information about the samples

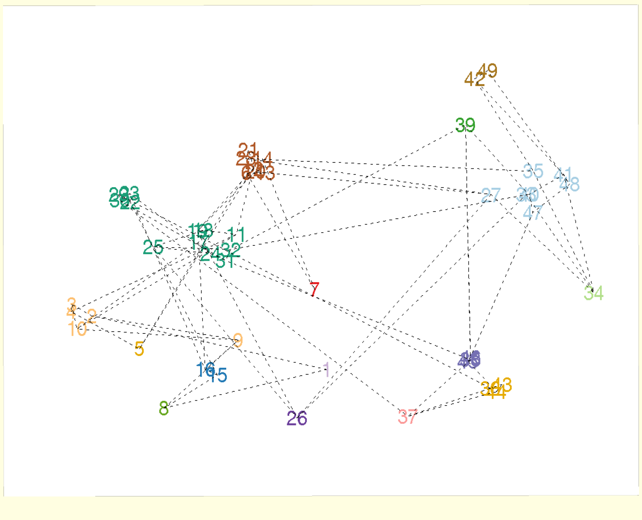


Conclusion

o Stress the proximities and the dissimilarities between species
o Represent the within-species variability and highlight potential new cryptic species
o Assign new samples to a species

Neighborhood structure on a 7x7 grid




42
49
39
21
14
43
35
41
48
27
47
11
25
17
24
32
31
7
34
3
4
2
10
5
9
16
15
1
8
26
37
13