CIIT Islamabad, April 2012

Mathematical Tools for Biology

Francine Diener
University of Nice Sophia-Antipolis (UNS)
Laboratoire UNS-C.N.R.S J-A Dieudonné

My Problem : How to Teach Maths to Biologists ?

· Determine what are the mathematical tools most often used by professional biologists

· Find the tools that may be explained to beginners

· Keep in mind the main objectifs : convince of the usefullness of mathematics in biology and of its accessibility when needed.

An example : Cluster Analysis (CA)

Cluster Analysis is a set of automatized methods used to sort individuals (plants, cells, animals, genes, ...) into groups, called clusters, such that the degree of association is strong between members of the same cluster and weak otherwise.

Many domains of application :

· Ecology : plants systemics, phylogenies

· Transcriptomics : Express Sequence Tag (EST) or DNA Microarrays

· Evolutionnary Biology

· Medecine : sort different type of tissues, blood, etc... on medical images

How to form, automatically, the good clusters ?

· The data, n individuals for which one has p measurements , can be seen as a cloud of n points in R^p.

· To find the best partition of this set of points, it is impossible to look at all partitions and find the best one : there are far too many !

· The best we can do is to find an as good as possible partition by an iterative method.

· There exist many algorithms. We will present now two of them : **Hierarchical Cluster Analysis** and K-**means method**.

Hierarchical Cluster Analysis

Whatever the algorithm you have chosen, you have first to define a **distance** (or more precisely a dissimilarity) **between points** of the data set and **between clusters**.

The HA goes through a succession of partitions :

$- - \cdot$ First partition : n clusters of only one individual

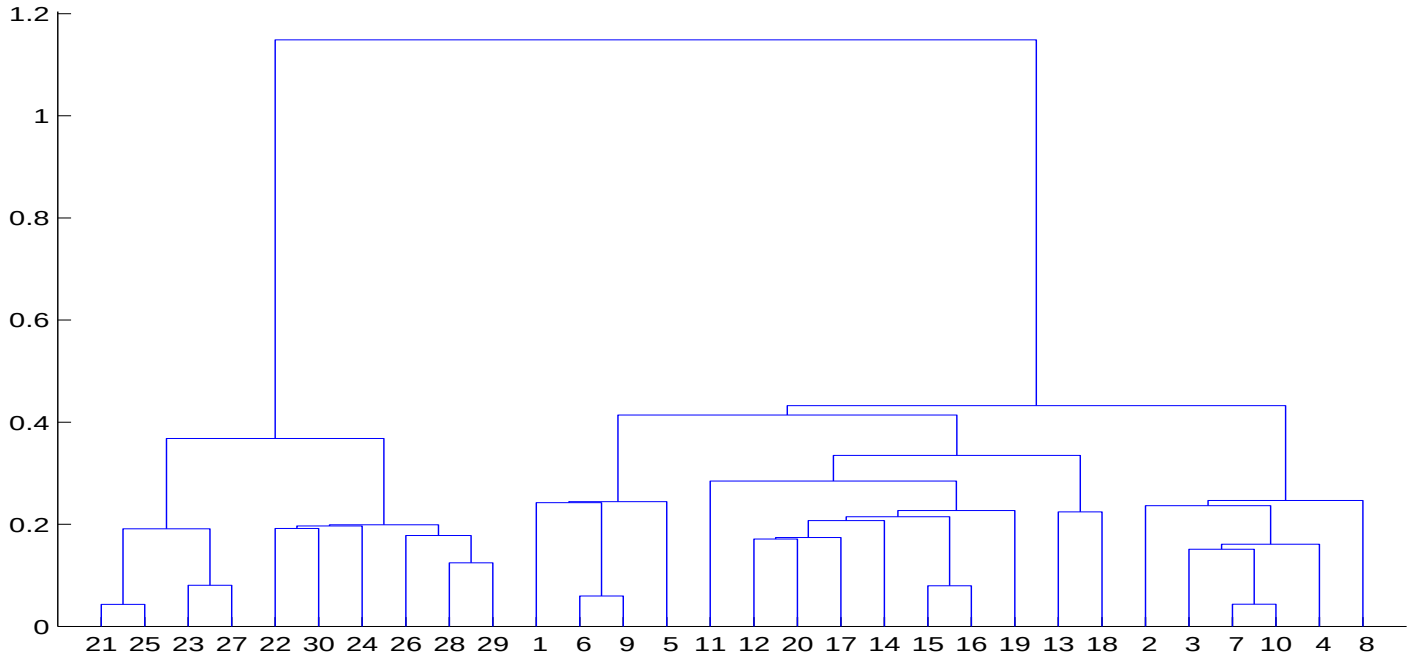
$- - \cdot$ Second partition : $n - 1$ clusters, after agregation of the two nearest points in a cluster.

$- - \cdot$ ....

$- - \cdot$ last partition : 1 cluster

Visualization : the dendrogram

A good way to summarize an Hierachical clustering is to plot a dendrogram.



It is also with the dendrogram that one chooses the "best" partition, by cutting it at the most appropriate hight.

K-Means Method

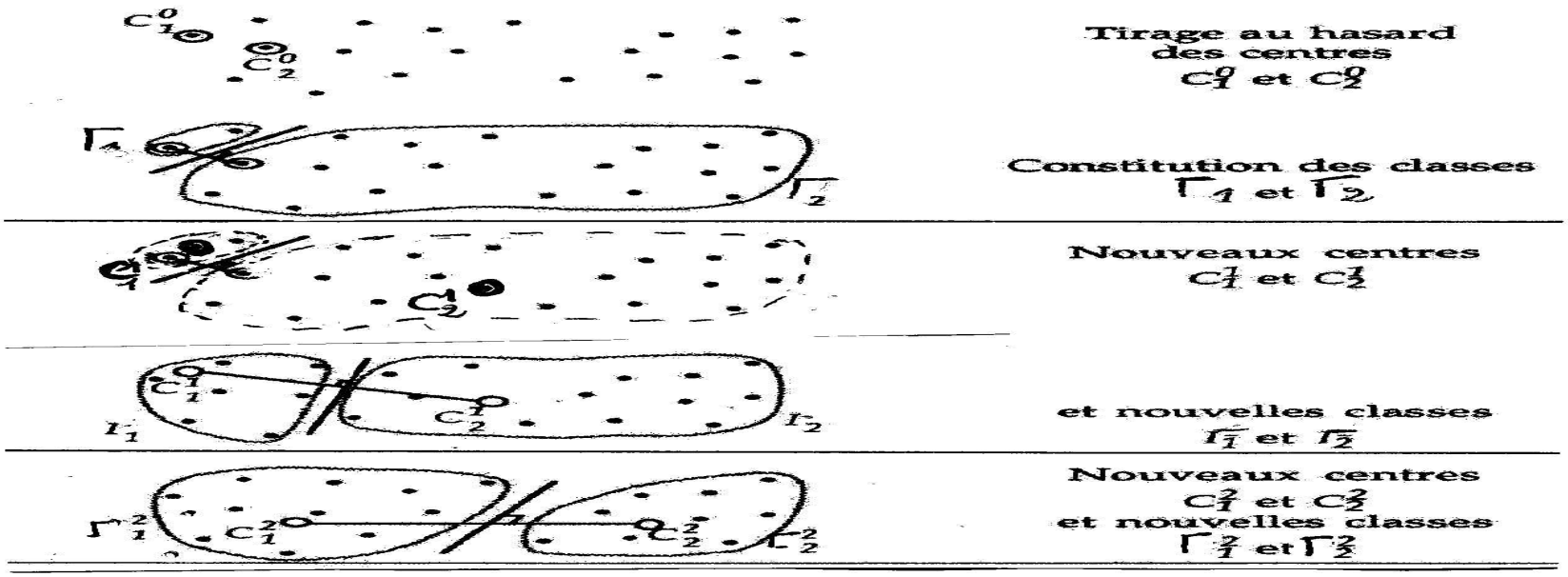
In addition to the distance between individuals and sets of individuals, one needs to choose in advance, for this method, **the number q of clusters**.

The K-means method goes through a succession of "means" also called *centers :*

$- \cdot -$ First set of q centers : C_1^1, C_2^1, ..., C_q^1 are randomly chosen among the individuals.

$- \cdot -$ Second set of centers : from the first set of centers, one builds a first partition, Γ_1^1, Γ_2^1, ..., Γ_q^1 by agregation of each individual to the closest center. Then each center of the second set of centers is defined as the centroïde (also called "mean") of the Γ_k^1, for $k = 1, .., q$.

K-Means method on an example



The algorithm is stopped either if an additional step induces no new modification of the $\Gamma's$ or the $C's$, or after a given number of steps.

How to teach this to first year students in Biology ?

Two hours per week (10 weeks each semester)

- 1 hour lecture using a 2 pages lecture notes - 1 hour for exercices on one **answer-sheet** (2-4 pages)

The students **work by themself**

They are encourage to **discuss in small groups**

The teacher collects the **answer-sheets** each week and mark them

A example of filled in version of each **answer-sheet** is made available on the web with the lecture notes

How to prepare LN + AS ?

For the Lecture Notes :

we need to learn and well understand the mathematical tools and its use in Biology

we have to adopt a written style much "light" than usual (based on concret examples)

For the Answer-sheets :

we need to find realistic biological examples

we need to aks only questions easy to answer for the student himself

we have to look at all answer-sheets each weeks (and write a corrected version)

Why does the K-Means method work ?

Proposition 1 *For any partition of a set* Γ *into* q *clusters* Γ_1, ...,Γ_q, *the total inertia* $\mathcal{I}(\Gamma)$ *is the sum of its* **within-clusters inertia** *and its* **between-clusters inertia :**

$$\mathcal{I}(\Gamma) = \mathcal{I}_{within}(\Gamma) + \mathcal{I}_{between}(\Gamma)$$

where $\mathcal{I}_{within}(\Gamma) = \mathcal{I}(\Gamma_1) + \ldots + \mathcal{I}(\Gamma_q)$ *and* $\mathcal{I}_{between}(\Gamma)$ *is the inertia of the weighted* $(\overline{x}_k, \pi_k)$, *where* $\overline{x}_k$ *is the centroïd of* Γ_k *and* π_k *its weight.*

Why does the K-Means Method work ..... usually ?

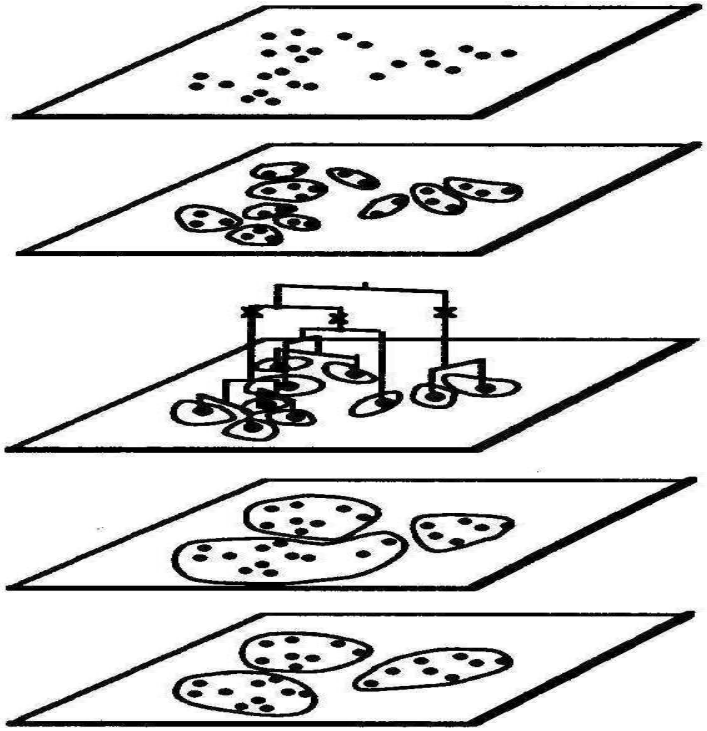
At each step, the K-means algorithm replace the previous partition of the data set Γ by another partition with a smaller within cluster inertia (or a larger between cluster inertia).

$$\mathcal{I}(\Gamma) = \mathcal{I}_{within}(\Gamma) + \mathcal{I}_{between}(\Gamma)$$

As the number of partitions of Γ is finite, **this algorithm does converge**.

BUT : it may happens that it converges to some **local optimum** that could be highly suboptimal when compared to the global optimum !

Others methods



Données
avant la classification

1. Partition préliminaire :
- centres mobiles
- groupements stables

2. Classification ascendante
hiérarchique sur les centres

3 a. Partition finale en 3 classes
par coupure de l'arbre

3 b. "Consolidation"
par réaffectation

Others : dynamic K-means method, neuronal methods (Voronoi cells), etc...

Statistical learning