About the Mapper algorithm and its most famous application

Ruth Lang Fuentes

1 Introduction

The aim of topological data analysis is primarily to somehow get to know the shape of a set of data points lying in some metric space, like $\mathbb{R}^N$ for N very big. There have been developed several methods in the last decade to solve this problem. Persistent homology for example analyses the homological signatures of point clouds to understand their geometry. Another approach is to try to visualize the data or to find images attached to point cloud data to obtain a qualitative understanding of the data through direct visualization via for example a graph. Therefor Gurjeet Singh, Facundo Mémoli and Gunnar Carlsson developed in 2007 the *Mapper* algorithm described in detail in [SMC07]. Together with a so-called filter function *Mapper* produces (in its simplest version) a graph and thus manages to obtain a one-dimensional picture of the data that yet provides information on their topology. The key idea of *Mapper* is to identify local clusters within the point cloud and understand the interaction of these partial clusters with each other.
In the first section this algorithm will be described in order to then in the second section explain, how it was applied to breast cancer microarray gene expression data in 2010 by Monica Nicolau, Arnold J. Levina and Gunnar Carlsson ([NLC11]) and how this application of *Mapper* identified a subgroup of breast cancers that could not be found by just doing *Clustering*.

2 Mapper

We are interested in finding topological properties of a topological space in a combinatorial way, looking for example at an associated simplicial complex.

Definition 1 *Given a finite covering $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ of a space X, we define the nerve of the covering U to be the simplicial complex $\mathcal{N}(\mathcal{U})$ whose vertex set is the indexing set A, and where a family $\{\alpha_0, \alpha_1, ..., \alpha_k\}$ spans a k-simplex in $\mathcal{N}(\mathcal{U})$ if and only if $U_{\alpha_0} \cap U_{\alpha_1} \cap ... \cap U_{\alpha_k} \neq \emptyset$.*

The *Mapper* construction is motivated by the following topological construction that associates a simplicial complex to a topological space via a continuous function, improving therefore the construction of taking the nerve of some covering:

Let X, Z be a topological spaces, $f : X \to Z$ a continuous function. From an finite open covering $\mathcal{U} := \{U_\alpha\}_{\alpha \in A}$ of Z we get, since f is continuous, an finite open covering of X, namely $\{f^{-1}(U_\alpha)\}_{\alpha \in A}$. For each α consider now the decomposition of $f^{-1}(U_\alpha)$ into its path connected components and let $\overline{\mathcal{U}}$ denote the covering of X by this path

connected components obtained from the covering $\mathcal{U}$ of Z. Its nerve $\mathcal{N}(\overline{\mathcal{U}})$ can now be used to provide information about the topological space X.

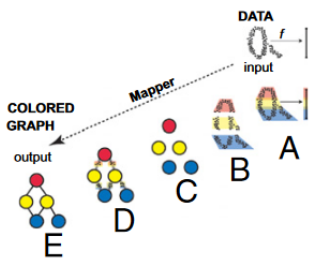
Let now for the whole section X be a **finite point cloud**, and therefore discrete, endowed with a metric d. The idea of *Mapper* is to now describe a method that transports this construction from the setting of topological spaces to the setting of point clouds, where the notion of "connected components" does not make sense as we work with a discrete set. It is replaced by *clustering*, which turns out to be the appropiate analogue. *Mapper* does not place any conditions on the clustering algorithm, however the main example of such an algorithm is the so-called *single linkage clustering*:

Definition 2 *Fixing the value of a parameter $\epsilon > 0$ one defines single linkage clustering, the clustering algorithm that is used for Mapper, as follows: Two points $x, x\prime \in X$ belong to the same cluster if and only if the they belong to the same equivalence class of the equivalence relation generated by $\backsim_\epsilon$ defined by $x \backsim_\epsilon x\prime :\Leftrightarrow d(x, x\prime) \leq \epsilon$.*

Definition 3 *The **Mapper algorithm** can now be defined, it can be implemented in five steps:*

1. *Define the so-called filter functions, continuous functions $f : X \to Z$ into a reference metric space Z which mainly is chosen to be $Z = \mathbb{R}$*
2. *Select a covering $\mathcal{U}$ of Z*
3. *From the covering $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ contruct the subsets $X_\alpha := f^{-1}(U_\alpha) \subseteq X$*
4. *Select a value $\epsilon > 0$ and apply (single linkage) clustering with this parameter ϵ to the sets X_α to obtain the set of clusters. We now have a covering of X parametrized by pairs (α, c), where $\alpha \in A$ and c is one of the clusters of X_α*
5. *Construct the simplicial complex with vertex set consisting of the set of all possible such pairs (α, c) and where $\{(\alpha_0, c_0), ..., (\alpha_k, c_k)\}$ spans a k-simplex if and only if the clusters $c_0, ..., c_k$ have a point in common (cf. nerve of a covering).*

A simplier version of *Mapper*, which is the one applied to the data in section 2, aims at obtaining a graph. Therefor the covering $\mathcal{U}$ of Z is chosen so that at most two open sets of $\mathcal{U}$ intersect. Alternatively, you can just reduce Step 5 of the algorithm to: Connect to clusters by an edge if and only if the corresponding clusters have points in common. This is actually the version of *Mapper* used in [NLC11].
The vertices of the obtained graph are often colored by the average value of the filter function to get a more precise visualisation.
The following sketch visualizes the various steps of *Mapper* (the filter function used here is just a projection/ height function):



Remark 1 *The implementation of Mapper depends on a lot of choices, like the filter function f, the covering $\mathcal{U}$, the distances notion d in X, the parameter ϵ or the selected clustering algorithm.*

G.Singh, F. Mémoli and G. Carlsson give some examples where *Mapper* was applied. But it seems that one of the most significant applications of the algorithm was the one described in [NLC11] in 2011 as it led to an important discovery in medicine which until then could not have been detected using *clustering*.

3 Application to breast cancer microarray gene expression data

In their paper *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival* M.Nicolau, A.J. Levine and G.Carlsson introduce a method that extracts information from high-dimensional and very sparse microarray data by using Progression Analysis of Disease (PAD) which is an application of Mapper to transcriptionally genomic data, which are first transformed by so-called Disease-Specific Genomic Analysis (DSGA). They found out that visualization of the data via a graph has an advantage over clustering, since when the method was applied to breast cancer transcriptional data they could identify a unique subgroup of so-called *Estrogen receptor positive* (ER^+) breast cancers, whose tumor cells have more Estrogen receptors than healthy cells and therefore grow faster with Estrogen. This subgroup of ER^+ breast cancers has 100 % overall survival and no metastasis and forms a new subtype of breast cancer that cluster analysis is unable to detect. It was called $c-MYB^+$ breast cancer due to its high levels of expression of the $c-MYB$ gene.

3.1 Preliminary Mathematical tools

The mathematical method introduced in [NLC11] to unravel the geometry of the data sets is the so-called *Progression Analysis of Disease* (PAD). It is an application of the simple version of *Mapper* defined above to DSGA- transformed data. We will first define what DSGA- tranformation means:

Definition 4 *Disease-Specific Genomic Analysis (DSGA) transforms data from diseased tissue as a sum of a normal component $Nc.\overrightarrow{T}$, that mimics healthy tissue (obtained by computing a Healthy State Model) and a disease component $Dc.\overrightarrow{T}$, that measures error or deviation from normal and are the data used later when applying Mapper. Let $\overrightarrow{T}$ be the original tumor vector, then:*

$$\overrightarrow{T} = Nc.\overrightarrow{T} + Dc.\overrightarrow{T} \tag{1}$$

Definition 5 *The input we want to analyze is a data matrix from diseased tissue, in which the columns are patients and the rows are any genomic variable type, e.g. transcriptional microarray data, like the data we will apply the algorithm to later.*
Progression Analysis of Disease (PAD) can now be implemented in four steps:

1. *DSGA-transform the data: Obtain a concatenated matrix out of the matrix* Dc.mat *whose columns consist of the disease components of the original tumor vectors and the matrix* L1.mat *whose columns estimate the disease component of normal tissue (leave-one-out estimates of the deviation from healthy state by normal tissue data).*

2. *Threshold the data coordinates so that only the genes showing significant deviation from healthy state are retained the data matrix (with any appropriate test for significance).*

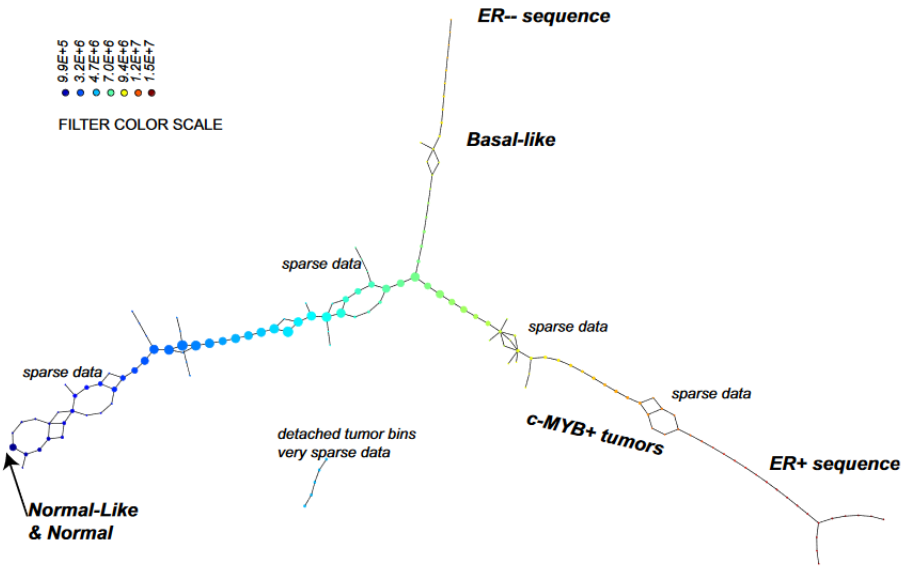
3. *Define Mapper filter function for the columns of the DSGA-transformed data matrix whose coordinates are the individual genes g_i. Let $\overrightarrow{V} = (g_1, ..., g_s)^t$ be such a data point. Then the filter function $f_{p,k}$ used is the vector magnitude in the L^p-norm, as well as k powers of this magnitude:*

$$f_{p,k}(\overrightarrow{V}) = (\sum |g_i|^p)^{\frac{k}{p}} \tag{2}$$

4. *Apply Mapper to the data obtained in step 2 using the filter functions $f_{p,k}$. The distance function used on the data is the correlation distance.*

3.2 Application and Interpretation

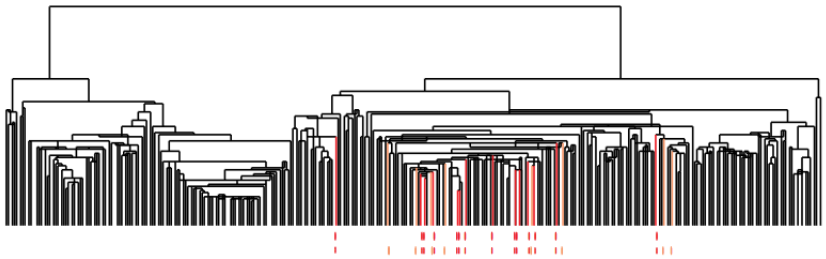
The steps of PAD were applied to a breast cancer microarray gene expression data set. The data consist of vectors in $\mathbb{R}^N$ which represent numerically the levels of gene expression in the respective genome. Each coordinate belongs to a specific gene in the genome of the tumor cell and describes how much mRNA is produced from the respective gene. Step 1 and Step 2 produced a data matrix with 262 rows, which describe the expression of the relevant genes. The values of the *Mapper* filter functions were then computed for $p = 1, ..., 5$ and $k = 1, ..., 10$. Several graphs were thus obtained which had to be then interpreted:
The **figure below** shows the resulting graph for $p = 2$ and $k = 4$. The local tumor cell clusters have been colored with repect to their average value the filter function, i.e. the color describes how much the gene expression of the respective diseased tissue deviates from healthy gene expression. If the cluster is colored in blue, this means that we got a low value of the filter function and thus gene expression is close to normal, whereas red stands for a large value of the filter function and therefore a large deviation from normal along multiple genes. This means, that many genes exhibit either increased or decreased activity relative to normal. Sparse regions in the data tend to form loops in the graph. The graph is composed of three branches representing different types of breast cancers: The ER^--sequence, the ER^+-sequence and the *Normal-Like* tissue, which is a subtype of ER^+-breast cancer. While the *Normal-Like* breast cancer subgroup was already known, a new unknown subgroup could now be identified by obtaining this graph via *Mapper*: A denser region of tumor clusters within the ER^+-sequence which is flanked by areas of sparse data. Due to its high levels of gene expression of special genes like $c - MYB$ it was denoted the $c - MYB^+$ group. There is furthermore low activity in other gene groups like innate inflammatory genes, relative to normal tissue. This extreme deviation from normal molecular profiles and the 100% overall survival, that was found out after doing PAD, suggest that the tumors of this new group have a mechanism to respond in a protective way.



Further, several biological analyses were done to the data to give evidence to the point that $c-MYB^+$ breast cancer is really a unique group, that does not fit into previously identified breast cancer types, and warrants being identified as a breast cancer group:
As already mentioned above *Survival analysis* found out that the $c-MYB^+$ group showed 100 % overall survival with no recurrence and no death from disease. This information was not incorporated earlier in PAD. *Molecular Subtype Classification* analyzed that the subgroup is not of one special molecular subtype. *Prediction Analysis of Microarrays* revealed that the subgroup is distinct from normal tissue and the *Normal-like* group and signicantly homogeneous as a class. Two predictor genes were able to distinguish between the $c-MYB^+$ group and normal tissue with error $= 0$. *Significance of the Analysis of Microarrays* detected a set of genes that are signicantly different between the $c-MYB^+$ group and normal samples or the $c-MYB^+$ group and the rest of the ER^+ sequence. Furthermore, *Cluster Analysis* was also applied to the data set:

3.3 Comparison with Cluster Analysis Applied to the same data matrix

Unlike the *Normal-like* tumor group, the new $c-MYB+$ tumor group is not visible when just doing cluster analysis. The outputs of (average linkage) clustering were compared to the *Mapper* results applied to the same exact DSGA-transformed data matrix to show that the two procedures are different. The $c-MYB+$ tumors (marked in red/ orange in the cluster dendrogram below) are scattered among different clusters, while PAD had been able to extract this group that turned out to be statistically and biologically coherent:



4 Conclusion

The $c - MYB+$ breast cancer group is a unique group that does not fit into previously identifed breast cancer types, shows uniformity in molecular signature and clinical and survival properties. It has also been validated in other breast cancer data sets.
The application and its result illustrate that TDA is particularly appropiate for the analysis of biological data. The viewpoint that it provides of these data is combinatorial and thus easy to grasp. Also TDA has a degree of robustness to the sort of distortions that can occur when studying biomedical data. Methods like *Mapper* are more sensitive than cluster analysis in identifying the subtle geometry of (genomic) data and can therefore uncover subgroups of diseases like the $c - MYB^+$ breast cancer group.
As a remark one can note, that many choices, that are made when applying *Mapper* to the data, seem to be kind of random and are not explained in [NLC11]. Nevertheless they obtained an important result for cancer research by means of *Mapper*.

References

[Car09] Carlsson, Gunnar: Topology and data. In: *Bulletin of the American Mathematical Society* 46 (2009), Nr. 2, S. 255–308

[NLC11] Nicolau, Monica ; Levine, Arnold J. ; Carlsson, Gunnar: Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. In: *Proceedings of the National Academy of Sciences* 108 (2011), Nr. 17, S. 7265–7270

[SMC07] Singh, Gurjeet ; Mémoli, Facundo ; Carlsson, Gunnar E.: Topological methods for the analysis of high dimensional data sets and 3d object recognition. In: *SPBG*, 2007, S. 91–100