Digital Signal Processing 48 (2016) 1–11

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

PKOM: A tool for clustering, analysis and comparison of big chemical collections

Christophe Molina *, Olivier Stepien, Bernard Pessegue, Jean-Philippe Rameau

Structure, Design and Informatics, Lead Generation & Candidate Realization Department, Sanofi, 195, route d'Espagne, 31036 Toulouse Cedex, France

ARTICLE INFO

Article history:
Available online 11 September 2015

Keywords:
Self-organizing maps
Tree maps
Chemoinformatics
Clustering
Organization of chemical collections

ABSTRACT

We describe the algorithm underlying PKOM, a methodology for clustering, analysis and visualization of multi-dimensional data onto a two-dimensional map. PKOM is based on the mixture of two very popular methods that have been widely used by the pharmaceutical industry for the clustering of genomic or SAR (Structure Activity Relationship) chemical information. The first method at the origin of PKOM is SOM (Self-Organizing Maps), a clustering technique based on neural networks. The second method is TREE MAPS, a visualization method based on hierarchical clustering by dendrograms. We initially describe herein the two methods and the reasons why we have taken the best of both to merge them into PKOM. We then describe in detail the PKOM algorithm and its advantages compared to the two former. Examples are given on how to apply this kind of 2-D topological clustering technique to the organization of big pharmaceutical collections in practical cases.

1. Introduction

The advent of combinatorial chemistry and the evolution of pharmaceutical chemical estates to huge medicinal chemistry collections have forced chemoinformatics to develop new tools to topologically cluster, visualize and manipulate such kind of data. A first attempt started with the use of Neural Networks techniques [3] such as SOM [4] and was followed by hierarchical clustering techniques such as TREE MAPS [2], which both facilitate not only the organization of the data but its visualization as well.

Self Organizing Maps (SOM, known too as KOHONEN maps from the author's name) is a class of machine-learning algorithm based on unsupervised training [1,4]. It belongs hence to a kind of methods widely known under the category of Neural Networks [5]. As the name indicates, these techniques build around the so-called Neurons (a mathematical equation that often reduces to a simple nonlinear mathematical equation per neural cell) connected as a Network that allows either the classification or clustering of data. In the case of SOM, every neuron is a vector representing the center of gravity of a cluster of objects in a high dimensional space, whatever the object and the way it is mathematically represented (see Figs. 1a and 1b).

If no topology is added to the SOM algorithm, a KOHONEN map is identical to the K-Means algorithm [6], which trains separately a predefined number of centers of gravity based on data, starting from random centers until they occupy the underlying centers of the real data (see Fig. 2a). However, in most of the cases, SOM are trained to eventually produce a topology that intends to translate the organization of the data from the high dimensional space onto a two-dimensional grid. This is achieved following the individual training of neural cell centroids, through first randomly placing the neural cells onto this grid and then averaging the centers of gravity that are within the same neighborhood in the grid. The alternation of these two steps – the calculus of centers of gravity by K-Means and the averaging of them in their vicinity of the map, leads to a two dimensional topological representation where the data that are in the same cell are similar (as in Fig. 1b) and the cells that are in the same neighborhood of the grid contain clusters of data nearby in the D-dimensional space (see Fig. 2a and 2b).

At a first glance, this algorithm seems interesting but a more in depth analysis would show that the two steps are in conflict – what the K-Means step builds as neural centers is faded by the average step and what the averaging of centers achieves is again disturbed by the following K-Means step. Computational performance is hence low and the algorithm would never reach convergence if the averaging would not be controlled by a penalizing parameter at every training cycle that reduces the averaging neighborhood radius to zero, which is in the end the K-Means algorithm.

* Corresponding author.
E-mail address: christophe.molina@pikairos.com (C. Molina).

http://dx.doi.org/10.1016/j.dsp.2015.08.010
1051-2004/

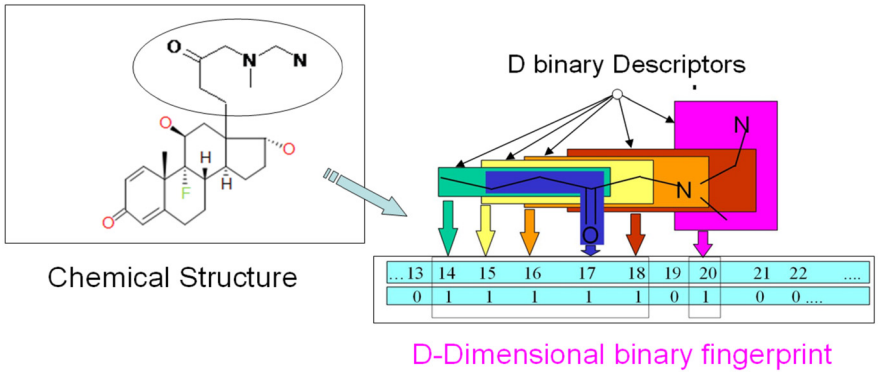

Fig. 1a. Chemical structures are usually fragmented to generate structural descriptors. The figure illustrates a linear fragmentation of a portion of a structure (within the ellipse) that is translated into the presence or absence of a set of binary descriptors referenced by an index. The whole fragmentation of a compound in this manner ends up with the generation of a binary fingerprint, made in total of D binary descriptors. Once the chemical structure is fragmented, it can be mathematically represented as a point in a D-dimensional space.

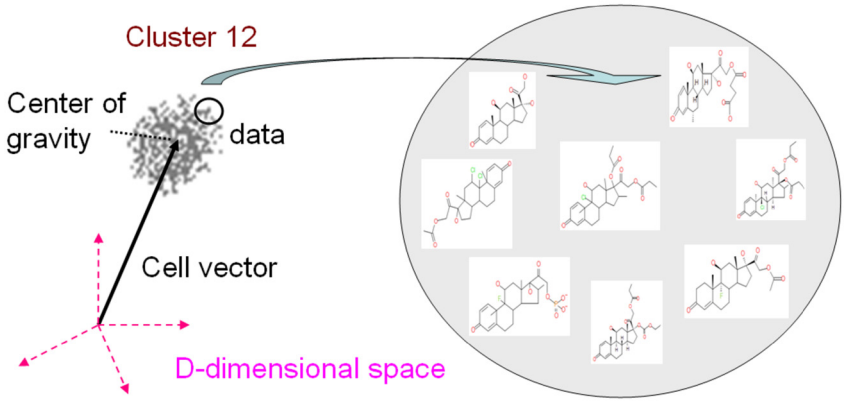

Fig. 1b. Chemical structures described by a fingerprint made of descriptors are usually represented in a D-dimensional space. Similar chemical structures gather together into clusters with a center of gravity, represented by a vector with respect to its D-dimensional coordinates.

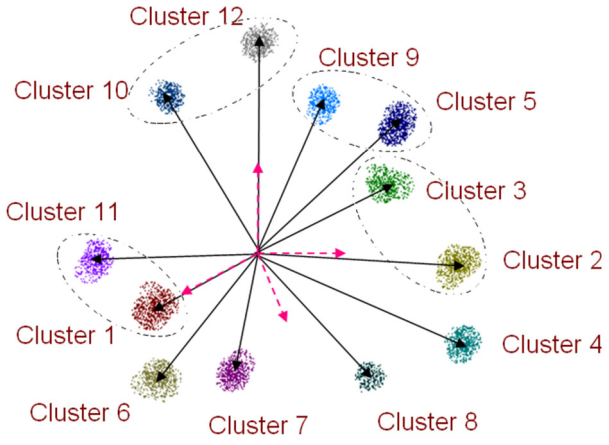

Fig. 2a. A chemical collection is usually made of series of chemical structures that gather together by similarity into clusters when represented into a high D-dimensional space (illustrated here by magenta coordinates.) Every cluster has its own center of gravity attached to a vector in the D-dimensional space. Clusters that lay nearby in this space are represented by centroids close each other and hence by vectors pointing to loci nearby in the space. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

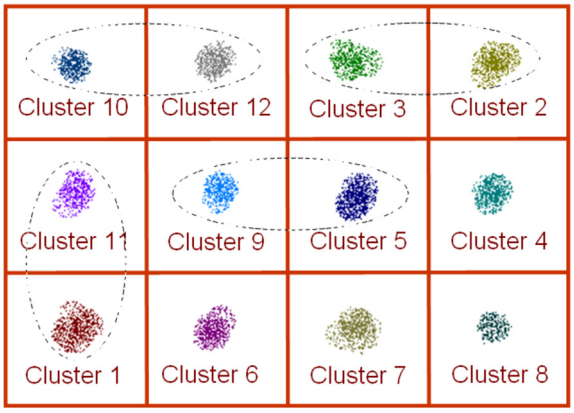

Fig. 2b. The 12 clusters depicted in Fig. 2a are schematically organized onto a Self Organizing Map (SOM), a 2-D topological representation that tries to preserve the underlying high-dimensional vicinity of clusters in a 2-dimensional grid. Every cell of the map hosts a cluster represented by a center of gravity. Consequently every cell contains chemical structures that have been gathered based on similarity. Neighboring clusters, such as 10 & 12 or 1 & 11, end up as neighbors in the topological organization of a SOM, as schematically shown in this minimalist picture (see Fig. 6 for a real SOM).

Tree Maps are a hierarchical visualization technique based on a dendrogram, which allows the representation of high dimensional data onto a two-dimensional space [2]. Many different implementations have been published since the invention of Tree Maps, some based on a free structure representation on a 2-D topological space [7,9] and others based on a fixed 2-D grid [8,10], the same as SOM. The Tree Map algorithm is also made of two steps. The first one hierarchically clusters the data using a binary dendrogram [11] – a tree where every node is connected at most to two branches or leaves (Fig. 3).