Multivariate analysis of toxicity experimental results of environmental endpoints

Kamel Mansouri, Matt Martin and Richard Judson

National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, RTP, NC, USA.

Kamel Mansouri | mansouri.kamel@epa.gov | 919-541-0545

Introduction

The toxicity of thousands of chemicals has been assessed experimentally during the last decades in the framework of several projects of the EPA and its partners. A valuable list of over 300 experimental studies of animal toxicity testing on a dataset of hundreds of chemicals were collected in ToxRefDB. In addition, ToxCast project released about a 1000 of High-Throughput Screening (HTS) bioassays for a large number of chemicals on 3 multi-year phases.

In order to optimize the use of these billions of dollars worth *in-vivo* and *in-vitro* studies in the toxicity assessment of chemicals, a more understanding of the information encoded in these databases is needed. In this work, different multivariate analysis techniques were employed to uncover the relationships between the different layers of information represented by the chemicals, the bioassays and the biological targets.

The results provided by these methods; Principal Component Analysis (PCA), Kohonen-Networks Self Organizing Maps (SOM) and Hierarchical clustering, were interpreted and used to map the linkage between chemicals and biological activity. This analysis aimed to determine the chemicals with high priority in the regulation process.

Materials and Methods

The Dataset:

The data considered for this study consisted of 524 chemicals tested in the guideline rat 2-year chronic cancer bioassay. There are a total of 156 endpoints included in the analysis. The 40 endpoints that showed no activity were removed before starting the analysis. The concentrations at which the chemical compounds showed activity were reported. This dataset was also converted into binary bioactivity fingerprints by generating 15 rows for each chemical at different doses.

Techniques employed:

Principal Component Analysis (PCA) and Multi-Dimensional Scaling (MDS):

PCA is a widely used tool for reducing dimensionality and for visually exploring the data. It maximizes the captured variance of the initial variables by projecting them into a lower number of new variables called Principal Components (PCs). MDS similarly reduces dimensionality, but conserves the distances between the samples in lower dimensionality for visualization purposes.

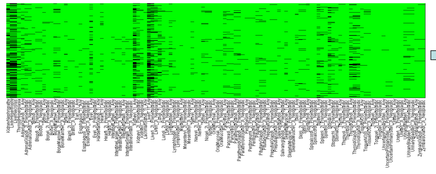
Self Organizing Maps (SOM):

The Kohonen SOM method was initially developed as an unsupervised technique for machine learning and clustering of data. It employs Artificial Neural Networks (ANN) to collapse the data samples in a predefined number of neurons in a two dimensional space. The winning neurons contain clusters of samples.

Hierarchical Clustering:

In this work, hierarchical clustering is implemented using different linkage algorithms (average, single, complete, Ward). Several metric distances such as the Euclidean, cosine and Jaccard-Tanimoto were applied for a better understanding of the data layout.

Fig1: Heat Map showing the activity of the chemicals on the different endpoints



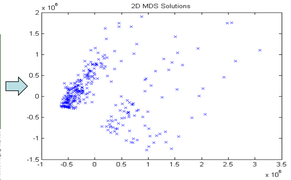


Fig2: 2nd order MDS of the dataset

Hierarchical clustering

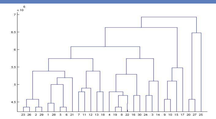


Fig3: Euclidean Complete linkage clustering

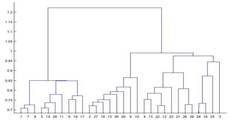


Fig4: Cosine Average linkage clustering

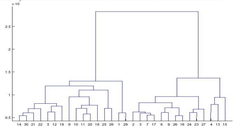


Fig5: Euclidean Ward linkage clustering

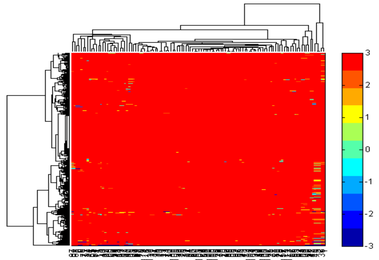


Fig6: Ward linkage clustered Heat Map using Euclidean distance on scaled dataset.
The colormap is ranging from blue to red representing actives and non-actives, respectively.

The clustering methods show different patterns and numbers of clusters. However, most of the clusters at the lower level are overlapping.

Self Organizing Maps

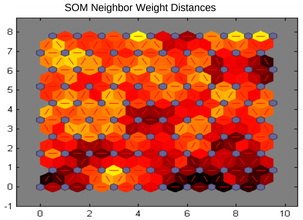


Fig7: Weights of the distances between the winning neurons showing the similarity of the clusters on the scaled original dataset.

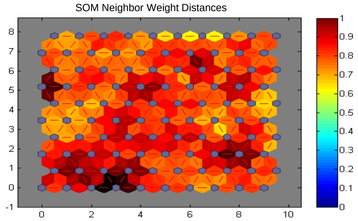


Fig9: Weights of the distances between the winning neurons showing the similarity of the clusters on the fingerprinted dataset.

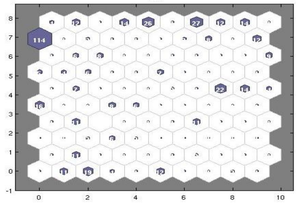


Fig8: The winning neurons showing the number of the chemicals in each cluster of the scaled original dataset.

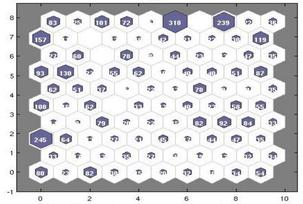


Fig10: The winning neurons showing the number of the chemicals in each cluster of the fingerprinted dataset.

Principal Component Analysis

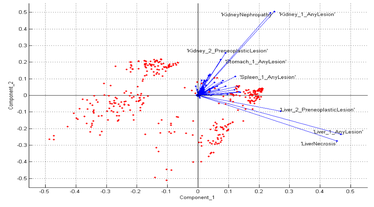


Fig11: PCA biplot showing the loadings of the assays and the scores of the chemicals projected on the 2 first principal components. PC1 captured 18.65% of the variance while PC2 captured 10.47% of the variance in the original scaled data matrix.

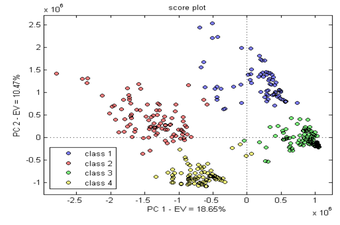


Fig12: K-means clustering of the original scaled data matrix projected into the first 2 PCA components showing 4 different clusters of chemicals.

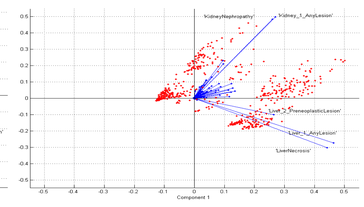


Fig13: PCA biplot showing the loadings of the assays and the scores of the chemicals projected on the 2 first principal components. PC1 captured 15.50% of the variance while PC2 captured 10.63% of the variance in the fingerprinted data matrix.

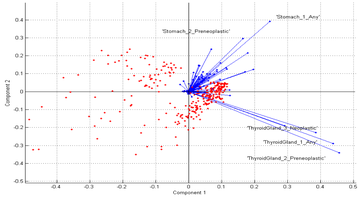


Fig14: PCA biplot showing the loadings and scores after removing the highest loadings assays of the first PCA. The first principal component PC1 captured 7.15% of the variance while PC2 captured only 5.85% of the variance in the original scaled data matrix.

The original scaled dataset and the fingerprinted data showed the same pattern of 4 clusters for the first two PCs that explained the maximum variance of the data. However, removing the most significant loadings resulted in a new layout and explained less of the variance for the first two components.

Conclusions and Further work

- The complex and unbalanced structure of this toxicity dataset required different multivariate analysis tools to better explore it and understand it.
  - The visualization tools showed the low number of active chemicals and an unbalanced distribution between the endpoints.
  - The clustering methods showed the diversity of this dataset.
  - The SOMs provided more detailed information about size of the different clusters and the distance (similarity) between them.
  - The fingerprint clustering associated different chemicals at different doses but with similar behavior among the endpoints.
  - PCA informed about the most significant assays in the data and their contribution to the clustering procedures.
- Further work:
  - Descriptor calculation and structural analysis of chemicals at the molecular level.
  - Re-clustering the data based on the molecular structures.
  - Investigate overlaps of the different clusters for Structure-Activity interpretation.

U.S. Environmental Protection Agency
Office of Research and Development