Chemistry Central Journal

ChemistryCentral

Oral presentation

Graph representation of molecular datasets: applications to dataset visualization and comparison using graph indices

A Tropsha* and D Fourches

Address: Laboratory for Molecular Modeling, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
* Corresponding author

from 4th German Conference on Chemoinformatics
Goslar, Germany. 9–11 November 2008

Published: 5 June 2009

Chemistry Central Journal 2009, **3**(Suppl 1):O9 doi:10.1186/1752-153X-3-S1-O9

This abstract is available from: http://www.journal.chemistrycentral.com/content/3/S1/O9

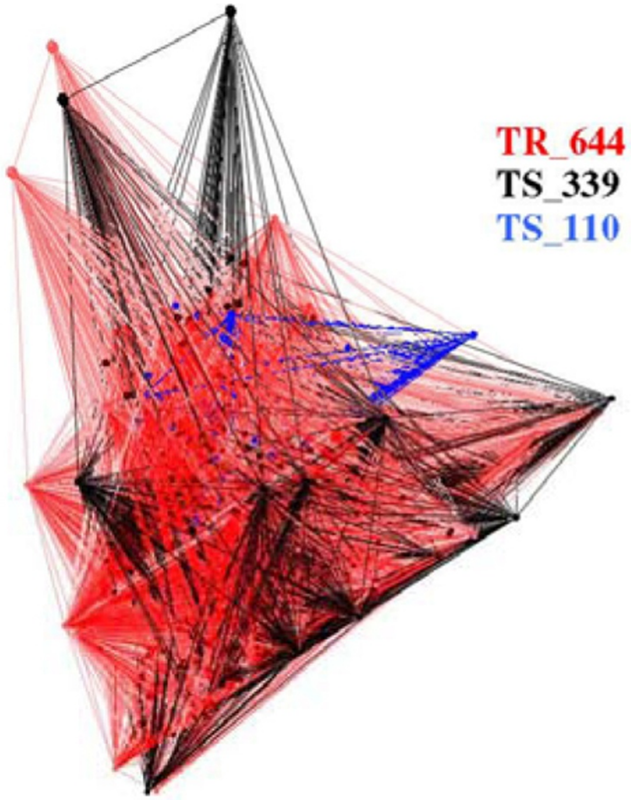

Figure 1
Advanced Dataset Graphs built for the "aquatic toxicity datasets, including the training set TR (red – 644 compounds), and two external sets TS, involving 339 (black) and 110 (blue) compounds respectively.

The visualization of large chemical datasets in multidimensional chemistry spaces is a great challenge in cheminformatics. To this end, we have developed a novel approach using graph representations for ensembles of molecules-points defined by their coordinates in high dimensional descriptor space. The Advanced Dataset Graph (ADG) consists of an ensemble of vertices (corresponding to molecules) connected by edges; two vertices are connected by an edge if the Euclidean distance between the vertex-molecules in the original descriptor space is within a user-defined cut-off. The ADDAGRA program has been implemented to build and visualize ADGs in three-dimensional graphic user interface. The uniqueness of this data representation is in that the points are projected onto 3D space using conventional Principal Component Analysis; however the projected vertices are connected by edges only if they are defined as neighbours in the original high dimensional space. Thus, unlike all other data projection approaches the ADG visualizes compound clusters in the original descriptor space exactly. In addition to the visualization, we have also implemented several simple graph indices for quantitative description and comparisons of ADGs. Three case studies involving (i) 101 AmpC beta-lactamase inhibitors, (ii) 2029 organic compounds with their measured intrinsic aqueous solubility, and (iii) 1093 chemical toxicants (see Figure 1) tested against Tetrahymena Pyriformis, have been analyzed with our ADG approach. Results suggest that some graph indices such as the average vertex degree or Randic connectivity index have the ability to discriminate similar vs. dissimilar pairs of datasets and address several other common issues in cheminformatics such as detection of outliers, shared regions in chemical and property space, etc. We suggest that ADGs could find a broad use for many cheminformatics applications.