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

Denis Fourches and Alexander Tropsha *

** Laboratory for Molecular Modeling, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill NC 27599, USA,* ***alex_tropsha@unc.edu***

The representation and visualization of large chemical datasets in multidimensional chemistry spaces is a great challenge in cheminformatics. To this end, we have developed a novel approach, the Advanced Dataset Graph Analysis (ADDAGRA), which uses graph representations for ensembles of molecules-points defined by their coordinates in high dimensional descriptor space. The dataset graph (DG) represents an ensemble of vertex-molecules connected by edges; the edge connects vertices that have the Euclidean distance between them in the original descriptor space within a user-defined cutoff. The ADDAGRA program with a graphical user interface was developed to build, visualize in 3D and real time, and analyze DGs. The uniqueness of this data representation is that the points are projected onto 3D space using conventional PCA; however, the edges are defined between neighbors in the original high-dimensional space. Thus, unlike all other data projection approaches the ADDAGRA 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 DGs. 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 ADDAGRA. 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, finding shared regions in chemical and property space, *etc*. We suggest that the ADDAGRA approach may find a broad application in cheminformatics.

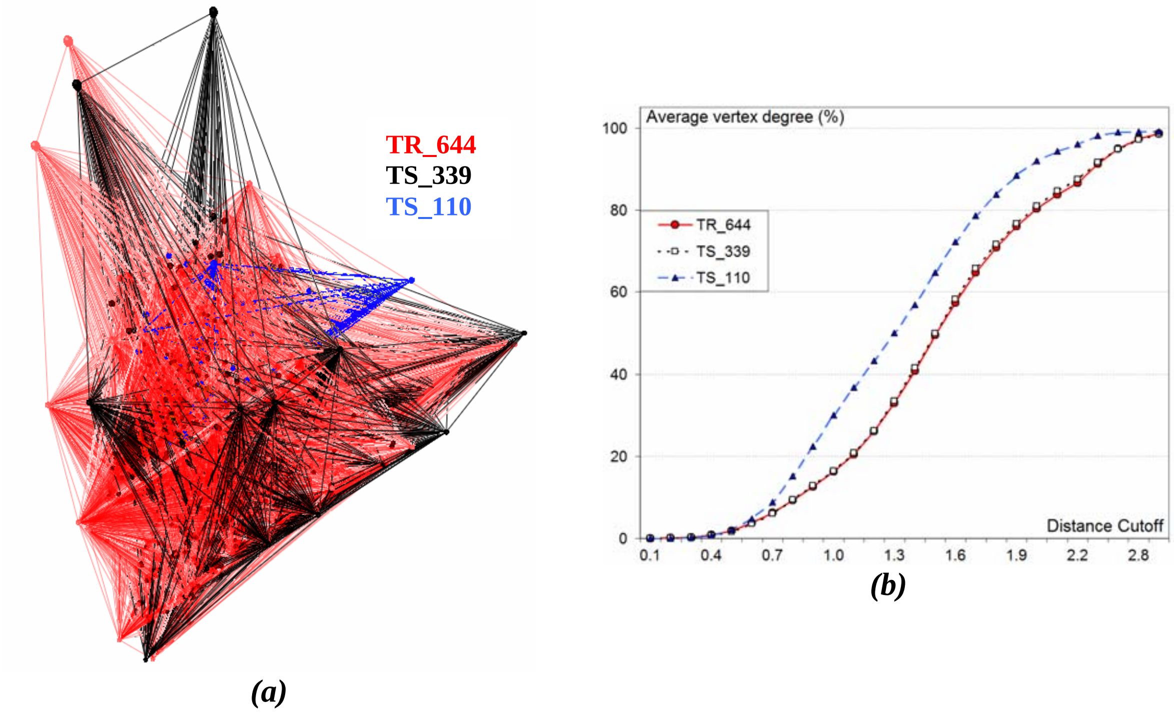

Figure 1. Application of ADDAGRA to the aquatic toxicity dataset. ***(a)*** Dataset Graphs for the training set TR including 644 compounds (*red*; TR_644), and two external test sets, including 339 (*black;* TS_339) and 110 (*blue;* TS_110) compounds respectively. ***(b)*** Average vertex degree as a function of the edge-defining distance cutoff for the three above datasets.