Journal of Chemical Information and Modeling 2015 vol.55 N1, pages 84-94

Chemical data visualization and analysis with incremental generative topographic mapping: Big data challenge

Gaspar H., Baskin I., Marcou G., Horvath D., Varnek A.
Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

© 2014 American Chemical Society. This paper is devoted to the analysis and visualization in 2-dimensional space of large data sets of millions of compounds using the incremental version of generative topographic mapping (iGTM). The iGTM algorithm implemented in the in-house ISIDA-GTM program was applied to a database of more than 2 million compounds combining data sets of 36 chemicals suppliers and the NCI collection, encoded either by MOE descriptors or by MACCS keys. Taking advantage of the probabilistic nature of GTM, several approaches to data analysis were proposed. The chemical space coverage was evaluated using the normalized Shannon entropy. Different views of the data (property landscapes) were obtained by mapping various physical and chemical properties (molecular weight, aqueous solubility, LogP, etc.) onto the iGTM map. The superposition of these views helped to identify the regions in the chemical space populated by compounds with desirable physicochemical profiles and the suppliers providing them. The data sets similarity in the latent space was assessed by applying several metrics (Euclidean distance, Tanimoto and Bhattacharyya coefficients) to data probability distributions based on cumulated responsibility vectors. As a complementary approach, data sets were compared by considering them as individual objects on a meta-GTM map, built on cumulated responsibility vectors or property landscapes produced with iGTM. We believe that the iGTM methodology described in this article represents a fast and reliable way to analyze and visualize large chemical databases.

http://dx.doi.org/10.1021/ci500575y