Mining Significant Chemical Substructures in 2D and 3D Spaces

Sayan Ranu and Ambuj K. Singh
Computer Science Department, University of California at Santa Barbara, Santa Barbara, CA 93106-9560
email: sayan@cs.ucsb.edu

Abstract: We present scalable and efficient techniques to mine molecular repositories based on their representation in the 2D and 3D spaces. Both techniques show immense potential in molecular classification and activity prediction. Empirical evaluation has confirmed the potential of these techniques.

Increased availability of large repositories of chemical compounds has created new challenges and opportunities for the application of data-mining techniques to problems in chemical informatics. Often chemical compounds are represented as graphs where the nodes represent atoms and the edges represent covalent bonds between them. Mining statistically significant subgraphs from such graph databases enable us to isolate over-represented molecular substructures in a database. These over-represented substructures are highly information rich and find application in classification, activity prediction, and scaffold hopping. Fig. 1a presents an outline of our technique. First, a representation of each graph is generated in the vector space by performing random walk with restarts. The histograms are next filtered to identify only those that correspond to significant substructures. Finally, these significant histograms are employed for activity prediction. The method performed better than well-known molecular fingerprints such as Daylight.

A common representation of molecules is also based on the geometric location of important features (such as donors, acceptors, hydrophobic cores, etc.) in 3D space. The underlying geometry of the pharmacophores is responsible for binding between compounds and targets as well as properties of compounds such as Blood Brain Barrier permeability. Existing pharmacophore based techniques are targeted towards extracting and optimizing a specific pharmacophore. Such an approach is limited in terms of the search space it can investigate in the drug discovery process. Often, multiple pharmacophoric targets need to be analyzed. We designed a new technique that eliminates the need to optimize pharmacophores against a specific target by defining a Joint Pharmacophore Space of chemical compounds, targets, and physicochemical/biological properties using the 3D geometry of pharmacophoric features. We mine this space directly to identify pharmacophoric patterns. Fig. 1b outlines our approach. The statistical significance of the mined patterns are substantiated by their low p-values. Further, these patterns demonstrate extremely promising performance in molecular classification by achieving superior prediction quality compared to molecular fingerprints and 3-point pharmacophores.