Abstract

Deciphering the activity of chemical molecules against a pathogenic organism is an essential task in drug discovery process. Virtual screening, in which few plausible molecules are selected from a large set for further processing using computational methods, has become an integral part and complements the expensive and time-consuming in vivo and in vitro experiments. To this end, it is essential to extract certain features from molecules which in the one hand are relevant to the biological activity under consideration, and on the other are suitable for designing fast and robust algorithms. The features/representations are derived either from physicochemical properties or their structures in numerical form and are known as descriptors.

In this work we develop two new molecular-fragment descriptors based on the critical analysis of existing descriptors. This development is primarily guided by the notion of coding degeneracy, and the ordering induced by the descriptor on the fragments. One of these descriptors is derived based on the simple graph representation of the molecule, and attempts to encode topological feature or the connectivity pattern in a hierarchical way without discriminating atom or bond types. Second descriptor extends the first one by weighing the atoms (vertices) in consideration with the bonding pattern, valence state and type of the atom.

Further, the usefulness of these indices is tested by ranking and classifying molecules in two previously studied large heterogeneous data sets with regard to their anti-tubercular and other bacterial activity. This is achieved by developing a scoring function based on clustering using these new descriptors. Clusters are obtained by ordering the descriptors of training set molecules, and identifying the regions which are (almost) exclusively *coming from active/inactive molecules. To test the activity of a new molecule, overlap of its descriptors in those cluster (interpolation) is weighted. Our results are found to be superior compared to previous studies: we obtained better classification performance by using only structural information while previous studies used both structural features and some physicochemical parameters. This makes our model simple, more interpretable and less vulnerable to statistical problems like chance correlation and over fitting. With focus on predictive modeling, we have carried out rigorous statistical validation.*

New descriptors utilize primarily the topological information in a hierarchical way. This can have significant implications in the design of new bioactive molecules (inverse QSAR, combinatorial library design) which is plagued by combinatorial explosion due to use of large number of descriptors. While the combinatorial generation of molecules with desirable properties is still a problem to be satisfactorily solved, our model has potential to reduce the number of degrees of freedom, thereby reducing the complexity.