Available online at www.sciencedirect.com

ScienceDirect

Discrete Applied Mathematics 154 (2006) 2484–2498

DISCRETE APPLIED MATHEMATICS

www.elsevier.com/locate/dam

An analysis of the redundancy of graph invariants used in chemoinformatics

Boris Hollas

Theoretische Informatik, Universität Ulm, D-89081 Ulm, Germany

Received 26 January 2005; received in revised form 29 March 2006; accepted 7 April 2006
Available online 24 May 2006

Abstract

Molecular descriptors play a decisive role for evaluating large virtual libraries and to predict biological or physicochemical properties of compounds. Topological indices are an important class of molecular descriptors, based on the graph of a molecule. A major problem is that many topological indices are considerably correlated, impeding data analysis and interpretation. Also, a size-dependent variance of topological indices adversely affects data processing by neural nets. Using random graphs as a model for molecules, we examine correlations and variance of an abstract topological index with independent vertex properties. We consider a random graph model making no assumptions on the distribution of graphs and a model on a fixed number of vertices in which edges are selected independently. We show that topological indices may be strongly correlated even for independent vertex properties. On the other hand, uncorrelated topological indices and indices with constant or $\Theta(1)$ variance can easily be obtained within the respective random graph models.

MSC: 05C80; 05C90; 92E10; 60C05; 68Rxx

Keywords: Molecular descriptor; Topological index; Random graph; Correlation

1. Introduction

Chemoinformatics is a discipline that emerged in the past 15 years to cope with rapidly rising costs for the development of pharmaceuticals and to provide methods to analyze large chemical data sets [20]. Computer-generated chemical libraries may easily contain over 10^8 compounds, about 10^{100} possible molecules are assumed to exist [28]. These amounts are far too large to be processed by even the most advanced laboratory methods such as high throughput screening, leaving the need for even faster methods. *Virtual screening* is a method to automatically evaluate huge libraries of compounds. Virtual screening is most effectively used to narrow down the choice of potentially interesting compounds among a set of structures. Virtual screening helps chemists to decide what libraries and compounds to synthesize, which structures to further examine, but also to analyze libraries of existing compounds. For the design of chemical plants, the reliable prediction of physicochemical properties saves the need for costly experiments.

All of these methods rely on a useful encoding of the information contained in a representation of a molecule. Of the large number of existing molecular descriptors [25,16], topological indices are among those having received most attention by mathematicians. *Topological indices* are graph invariants applied to the graph of a molecule [7,26,4]. Note

E-mail address: hollas@informatik.uni-ulm.de

0166-218X/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2006.04.002

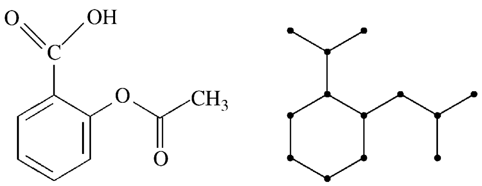

Fig. 1. Structure and graph of acetyl salicylic acid.

that a molecular graph may be labeled in an arbitrary way. Atoms, excluding hydrogen atoms, constitute the vertices of the molecular graph, covalent bonds between them form the edges (*H-depleted molecular graph*). Double or triple bonds are usually not taken into account, but may be allowed for by multiple edges. Fig. 1 shows the molecular structure of acetyl salicylic acid (Aspirin®) and its H-depleted graph.

Due to their minimal computational requirements, topological indices are well suited for the aforementioned virtual screening. On the other hand, we cannot expect them to be characteristic for a graph or molecule. It is believed that for any set of simple topological indices there may exist structures that will have identical values [24]. In practice, this is no big issue since chemist are more interested in relationships to molecular properties than in a high discrimination power of a topological index.

Many topological indices have the form

$$\sum_{u,v} x_u x_v, \tag{1}$$

where u, v are adjacent vertices or vertices at a distance $d \geqslant 1$ and x_u is some property of vertex u [25]. For example, the *autocorrelation index* is defined as

$$A_d = \sum_{\{\{u,v\}|d(u,v)=d\}} x_u x_v, \quad d \geqslant 0, \tag{2}$$

where x_v is a physicochemical property of atom (vertex) v [22]. If x_u is replaced by $\deg(u)$ in (2) the *Zagreb indices* are obtained [11,10].

Modifications of the autocorrelation indices to allow for the 3D structure of the molecule have been proposed [9]. As with all 3D descriptors, a limitation of 3D autocorrelation is that the final conformation of the molecule is often not known in advance.

Many topological indices exhibit considerable mutual correlation [19,27,2,3]. This is a major problem when performing structure-activity studies (SAR [5,15]) as the employed statistical methods may fail or give little meaningful results on sets of correlated data. In addition, strong correlations among a set of topological indices raise doubt whether these indices describe different and meaningful biological, chemical or physical properties of molecules [23].

Chemical similarity may be distorted and exaggerated if correlated descriptors are used in a similarity analysis.

If decision trees [6] are used in a SAR study, the trees are unstable, that is, completely different trees may be generated for similar data sets, making model interpretation difficult. If multivariate linear regression is used to derive a structure-activity relationship, the following may happen if descriptors are correlated too much:

- The regression is unstable and the p-values for the regression parameters β_i to be estimated not significant.
- The columns of the design matrix are linearly dependent, thus the least-squares estimator for $\boldsymbol{\beta}$ cannot be determined.
- Even worse, spurious estimates for the parameters β_i are produced due to the rounding error generated by the computer program used in “successfully inverting” the non-invertible design matrix [21].

Principal component analysis [1] can be applied to solve these problems, which, however, comes at the cost of model interpretability. In this case, principal components are linear combinations of all descriptors, which renders interpretation of the model, an important goal in SAR studies, impossible.