Evolution of Optimal Descriptors: Solved, Unsolved, and Unsoluble Tasks

Alla P. Toropova, Laboratory of Environmental Chemistry and Toxicology, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Milano, Italy
Andrey A. Toropov, Laboratory of Environmental Chemistry and Toxicology, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Milano, Italy

ABSTRACT

The quantitative structure - property / activity relationships (qsprs/qsars) analysis of different substances is an important area in mathematical and medicinal chemistry. The evolution and logic of optimal descriptors which are based on the monte carlo technique in the role of a tool of the qspr/qsar analysis is discussed. A group of examples of application of the optimal descriptors which are calculated with the coral software (http://www.insilico.eu/coral) for prediction of physicochemical and biochemical endpoints of potential therapeutical agents are presented. The perspectives and limitations of the optimal descriptors are listed. The attempt of the systematization of the models calculated with the coral software is the aim of this work.

KEYWORDS

Monte Carlo Method, Quasi-SMILES, SMILES, Super-Validation, Validation

INTRODUCTION

Simplicity provides clarity, reliability, and reproducibility, therefore aspiration to variety minimization was always peculiar to praxis of research work, in general, and for quantitative structure – property / activity relationships (QSPRs/QSARs) analyses, in particular (Montchamp et al. 1993). The optimal descriptor is an example of aspiration to the Simplicity.

The first attempts of the construction of the optimal descriptors were based on the molecular graphs (Randic & Basak 1999; Randic & Pompe 2001; Randic & Basak 2001; Da Silva Junkes et al. 2005). It is well-known that the majority of the topological indices are calculated with the numerical data on the vertex degrees (i.e number of vertices connected with the given vertex) in the molecular graphs, the numbers of paths of lengths 2, 3, etc., Morgan vertex degrees of the increasing orders: zero order is the usual number of neighbors. Extended connectivity of first order is calculated with the recurrence formula from data on usual vertex degrees, extended connectivity of second order is calculated with the recurrence formula from data on the extended connectivity of first order, and so on (Amic et al. 1998; Toropov & Toropova 2002a,b; Toropov & Toropova 2004).

The representation of the molecular structure by the molecular graph is a convenient way but in fact this representation involves the adjacency matrix (Randic & Basak 1999; Randic & Pompe 2001; Randic & Basak 2001), i.e., matrix with the nxn elements (n is the number of atoms in the molecule), whereas so-called simplified molecular input-line entry system (SMILES) (Weininger 1988; Weininger et al. 1989; Weininger 1990) is the representation by string of symbols. In other words, the representation by SMILES is more convenient and more “economical” than the representation by molecular graph for databases on the physicochemical and biochemical endpoints available on the
Internet. Under such circumstances, the optimal descriptors calculated from the SMILES (Toropov et al. 2008; Nesmerak et al. 2013; Toropova & Toropov 2014; Nesmerak et al. 2014; Masand et al. 2014; Veselinovic et al. 2014; Toropova et al. 2014) become an attractive alternative of the optimal descriptors which are calculated with the molecular graphs (Randic & Basak 1999; Randic & Pompe 2001; Randic & Basak 2001; Toropov & Toropova 2002a,b; Toropov & Toropova 2004).

However, the representation of the molecular structure by SMILES and the representation of the molecular structure by molecular graphs are different and moreover each of the above-mentioned representations has advantages as well as disadvantages. This leads to attractiveness of so-called hybrid optimal descriptors (Toropova et al. 2012; Achary 2014a,b; Fatemi & Malekzadeh 2015; Ghaedi 2015), calculated taking into account molecular attributes extracted from both the SMILES and molecular graphs.

During 2010s, the problem of creation of models for the endpoints relating to peptides and nanomaterials becomes important task of natural sciences (Toropov & Leszczynski 2006; Toropov et al. 2007). The representation of the molecular structure of nanomaterials (Toropov & Toropova 2015a,b; Toropova and Toropov 2015) as well as peptides (Toropov et al. 2012) is not convenient way to build up predictive models since these substances have very complex molecules. As an alternative of the traditional descriptors which are calculated with the graph or SMILES, the optimal descriptors which are translators of eclectic data represented by so-called quasi-SMILES (Toropov & Toropova 2015a; Toropova & Toropov 2015, Toropov et al. 2015) into endpoint prediction were developed to solve the task of building up models for endpoints related to peptides or nanomaterials (Toropova & Toropov 2013; Toropov & Toropova 2014; Toropov et al. 2015a,b; Toropova et al. 2015). Thus, the history of the optimal descriptors started by the modifications of the adjacency matrix and now comes to the phase of development of translation of available eclectic data into endpoints prediction.

The QSPR/QSAR analysis of drug-like substances (small organic molecules) is an important trend of mathematical and medicinal chemistry (Duchończ & Castro 2009; Deeb et al. 2011; Gozalbes at al. 2011; Castillo-Garit et al. 2012; Toropova & Toropov 2014; Castillo-Garit et al. 2015). The QSPR/QSAR analysis of peptides (large systems of amino acids) is also important and specific task of medicinal chemistry (Toropov et al. 2012; Almerico et al. 2013; Kandel et al. 2014; Chen et al. 2015). The estimation of ability of optimal descriptors to be a tool to solve the above tasks has been target of a few works (García et al. 2011; Garro Martínez et al. 2011; Mullen et al. 2011; Ibezim et al. 2012; Toropov et al. 2012; Veselinović et al. 2013; Li et al. 2014).

The aims of the present work are (i) the estimation of ability of the optimal descriptors which are calculated with traditional SMILES to be a tool to predict for potential therapeutic, physicochemical and toxicity endpoints; and (ii) the estimation of ability of the optimal descriptors calculated with quasi-SMILES which are representation of angiotensin-I-converting enzymes (ACE) to be a tool to predict ACE-inhibitory peptides activity (pIC50, n=234).

METHOD

Data: Substances examined in this work are drug like substances (http://www.uv.es/~galvez/tablevi.pdf). The mutagenic potential of these compounds has been analyzed in the literature (Kar & Roy 2011; Duchowicz et al. 2012; Toropova & Toropov 2014). Physicochemical data for a portion of these compounds are available on the Internet (U.S. National Library of medicine 2015). Four datasets have been organized using the database (I) normal boiling points (NBPoC, n=243); (II) octanol / water partition coefficient (logP, n=1451); (III) water solubility mg/L (logWS, n=879); and (IV) the lethal rat toxicity data in mg/kg: the predicted endpoint is the negative logarithm of 10000 of the lethal rate toxicity.
Related Content

Multilayer Perceptron Model for Predicting Acute Toxicity of Fungicides on Rats

QSAR-Models, Validation, and IIC-Paradox for Drug Toxicity
Artificial Neural Network (ANN) Modeling of Odor Threshold Property of Diverse Chemical Constituents of Black Tea and Coffee

Strategies of Virtual Screening in Medicinal Chemistry
www.igi-global.com/article/strategies-of-virtual-screening-in-medicinal-chemistry/191198?camid=4v1a