Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 25 (2006) 275–288
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Automatic extraction of structural alerts for predicting chromosome aberrations of organic compounds

Ernesto Estrada *, Enrique Molina

Complex Systems Research Group, RIAIDT, Edificio CACTUS, University of Santiago de Compostela, Santiago de Compostela 15782, Spain

Received 1 September 2005; accepted 8 January 2006
Available online 17 February 2006

Abstract

We use the topological sub-structural molecular design (TOPS-MODE) approach to formulate structural alert rules for chromosome aberration (CA) of organic compounds. First, a classification model was developed to group chemicals as active/inactive respect to CA. A procedure for extracting structural information from orthogonalized TOPS-MODE descriptors was then implemented. The contributions of bonds to CA in all the molecules studied were then generated using the orthogonalized classification model. Using this information we propose 22 structural alert rules which are ready to be implemented in expert systems for the automatic prediction of CA. They include, among others, structural alerts for *N*-nitroso compounds (ureas, urethanes, guanidines, triazines), nitro compounds (aromatic and heteroaromatic), alkyl esters or phosphoric acids, alkyl methanesulfonates, sulphonic acids and sulphonamides, epoxides, aromatic amines, azaphenanthrene hydrocarbons, etc. The chemico-biological analysis of some of the structural alerts found is also carried out showing the potential of TOPS-MODE as a knowledge generator.

Keywords: TOPS-MODE; Topological descriptors; Knowledge-generation; Clastogenicity; Chromosome aberration; Structure–toxicity relationships; QSAR

1. Introduction

In QSAR analysis a quantitative model is used to predict the biological response of a chemical based on a series of molecular descriptors or physicochemical properties [1]. However, the structural information contained in such descriptors or properties is encrypted [2] in a way that does not allow the extraction of structural rules to form a *knowledge base* similar to that provided by human expertise [3]. In the case of toxicological assessment of chemicals these knowledge bases are the heart of expert systems, such as DEREK [4] and TOPKAT [5], used to evaluate the toxicological profile of chemicals [6]. One of these toxicological endpoints which is of relevant importance is the chromosome aberration or clastogenicity produced by chemicals. Chromosome aberrations (CA) are DNA changes generated by different repair mechanisms of DNA double strand breaks, which are microscopically visible [7]. They are consequences of human exposure to ionising radiation or to mutagenic chemicals [8–11]. The frequencies of CA in peripheral lymphocytes show a positive correlation with the later onset of cancer in humans [7].

The necessity for the automatic generation of structural alerts for predicting CA and other toxicological endpoints is evident. On one hand, classical QSARs permit the classification of chemicals as clastogenic/nonclastogenic but their information cannot be easily incorporated on the existing expert systems due to the cryptic nature of the variables included in such models [2]. On the other hand, the traditional method for extracting knowledge from human expertise requires a great amount of (available) information about a set of chemicals permitting the expert their generalization. However, the rate of producing new chemical entities overtakes the rate of their toxicological profile evaluation. Thus a method that permits to extract knowledge from the minimum information available about a series of chemicals is necessary to keep expert systems updated. In this sense, an expert system can be considered as *knowledge archive* where a collection of knowledge is expressed using some formal representation language. An *automatic knowledge generator* is a methodology that will provide new structural alerts to the knowledge archive in a cyclic way keeping it updated. In previous works [12,13] we have shown that the so-called topological sub-structural molecular design (TOPS-MODE) approach [14–19] represents a useful platform for the automatic generation of toxicological structural alerts. In these works a general strategy for

* Corresponding author.
E-mail address: estrada66@yahoo.com (E. Estrada).

1093-3263/$ – see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2006.01.002

knowledge flow concerning skin sensitization based on the combined use of TOPS-MODE and DEREK expert system was proposed [12,13].

The main purpose of the current work is to generate structural alert rules that permit the identification of CA in chemical compounds using information coded in their molecular structure. Thus, we develop a classification model using the TOPS-MODE approach, which allows to calculate the contribution of each part of a molecule to the activity under study. Using this information we identify structural regions responsible for the clastogenic activity of chemicals and transform this information into structural alert rules which are ready to be implemented in expert systems such as DEREK.

2. Data set

A data set of 383 organic compounds compiled by Serra et al. was used for the purposes of the current study [20]. These compounds were selected among those reported on the *Compilation of Chromosomal Mutation Test Data* containing tests carried out by the *National Drug and Food safety Laboratory and the First Laboratory of the mutational Genetics Department of the Safety and Biotesting Research Center in Japan* [21]. These compounds were tested at two different times of exposure, mainly 24 and 48 h, in cultured Chinese hamster lung cells. After exposition, cells were processed by standard methods and chromosomal aberrations were identified. Compounds were classified as positive if there were 10% or greater aberrant cells and negative if there were 5% or less aberrant cells. Compounds classified as ''equivocal'' due to their percentage of aberrant cells (5–10% aberrant cells) were not included in this study as well as they were not considered in Serra et al.'s work. From this original data set three compounds could not be included in the current study as they have macromolecular structures, such as polymeric one (compounds 161, 185, 267 in Serra et al.'s work [20]). Three compounds in the original data set were salts of other compounds in the data set. For instance, compound 40 (in Serra et al. 's paper [20]) is aniline–HCl and compound 141 is aniline. Compound 11 is the sodium salt of the L-glutamic acid and compound 200 is L-glutamic acid. Finally, compound 324 is the salt of 115. In all cases salts were excluded from our data set. There are other five pairs of compounds which were geometric isomers distinguished neither by our approach nor by descriptors used by Serra et al. [20]. They are: 163/355, 42/136, 90/146, 348/361 and 166/175. In all cases one of the compounds in each pair was eliminated from our data set. Consequently, our data set is formed by 372 organic compounds including known carcinogens, drugs, food additives, agrochemicals, cosmetic materials, medicinal products, and household materials.

This data set was divided into two subsets, one containing 216 compounds (100 clastogenics and 116 nonclastogenic) was used as a training set for developing the classification model. The other formed by 156 compounds (11 clastogenic and 145 nonclastogenic) was used as a prediction set. Our main objective is to extract as much structural information as possible from this data set in order to formulate structural alerts for clastogenicity. Consequently, we keep the minimum number of clastogenic compounds out of the training set. In fact, we selected only those compounds used by Serra et al. [20] as the prediction set for the *k*-nearest neighbour model, i.e., we do not use any cross-validation set. In that work, however, the number of nonclastogenic compounds in the training sets is very much higher than the number of clastogenic ones. For instance, for the *k*-NN model development they used 245 nonclastogenic compounds and 101 clastogenic compounds and for SVM model development the training set consisted on 218 nonclastogenic and only 90 clastogenic compounds. Here we preferred to have a more compensated training set having approximately the same number of clastogenic and nonclastogenic compounds. Consequently, we selected at random several nonclastogenic compounds originally in the training set to be in the prediction set. This produced a training set having 116 nonclastogenic and 100 clastogenic compounds and the prediction set was finally conformed by these compounds plus those originally in the prediction set [20].

3. Methodology

3.1. The TOPS-MODE approach

In the last 10 years we have developed an approach to QSAR/QSPR based on the use of spectral moments of the bond matrix as molecular descriptors. It is known as TOPS-MODE approach, which is the acronym of topological substructural molecular descriptors/design [14–19]. TOPS-MODE approach is based on the calculation of spectral moments of molecular bond matrices appropriately weighted to account for hydrophobic, electronic and steric molecular features. Spectral moments are the trace of the *k*th power of a matrix, i.e., the sum of all the main diagonal entries of such matrices [14–16].

A bond matrix is a square symmetric matrix in which non-diagonal entries are ones or zeroes if the corresponding bonds have a common atom or not, respectively [22]. These matrices represent the molecular skeleton without taking into account hydrogen atoms. Bonds weights are placed as diagonal entries of such matrices and represent quantitative contributions to different physicochemical properties. Among bond weights currently in use in our approach we have standard bond distance (SD), standard bond dipole moments (DM), hydrophobicity (H) [23], polar surface area (PS) [24], polarizability (Pol) [25], molar refractivity (MR) [25], van der Waals radii (vdW) [26], and Gasteiger–Marsilli charges (Ch) [27].

The starting point for our approach is to calculate TOPS-MODE descriptors of the different types, e.g., H, PS, Pol, MR, vdW, and Ch, for the series of molecules under study. Then, we develop a quantitative model describing the property under study in term of the spectral moments. In general this model can be of the following form:

$$P = b_0 + \sum_{j=1}^{L} b_j \mu_j \quad (1)$$