European Journal of Medicinal Chemistry 46 (2011) 5910–5916

Contents lists available at SciVerse ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

Fragment-based QSAR model toward the selection of versatile anti-sarcoma leads

Alejandro Speck-Planche [a,*], Valeria V. Kleandrova [b], Feng Luan [a], M. Natália D.S. Cordeiro [a,*]

[a] Requimte/Department of Chemistry and Biochemistry, University of Porto, 4169-007 Porto, Portugal
[b] Faculty of Technology and Production Management, Moscow State University of Food Production, Volokolamskoe shosse 11, Moscow, Russia

ARTICLE INFO

Article history:
Received 6 April 2011
Received in revised form
24 September 2011
Accepted 29 September 2011
Available online 10 October 2011

Keywords:
Multi-target QSAR
Anti-sarcoma activity
Linear discriminant analysis
Fragments

ABSTRACT

A sarcoma is a type of cancer which is originated from the connective tissue cells. With the time, several sarcomas have become resistant to the current anti-tumoral drugs. Many works have been reported in order to explain some mechanisms of resistance in different types of sarcomas and around 2000 compounds have been tested as anti-sarcoma agents against several sarcoma cell lines. However, there is no an available methodology for the rational design of compounds with anti-sarcoma activity. The present work develops a unified fragment-based approach by employing a multi-target QSAR model for the efficient search and design of new anti-sarcoma agents against 12 sarcoma cell lines. The model was obtained with the use of a heterogeneous database of compounds and it was based on substructural descriptors. The percentages of correct classification of active and inactive compounds were higher than 85% in both cases. Also, the present approach provided the rapid extraction of substructural alerts responsible of anti-sarcoma profile by calculating the quantitative contributions of fragments to anti-sarcoma activity. To our knowledge, this is the first attempt to calculate the probabilities of anti-sarcoma activity of compounds against several sarcoma cell lines simultaneously, using a unified fragment-based QSAR model.

1. Introduction

Sarcomas are types of cancers that arise from transformed connective tissue cells. These cells can be originated from embryonic mesoderm, or middle layer, which forms the bones, cartilages, and fat tissues. Sarcomas can be considered as tumors of putative mesenchymal origin, which represent less than 1% of adult solid malignancy. However, the histological and biological spectrum of this disease is remarkable in comparison to the more prevalent epithelial-based malignancies [1]. Comprising a family of more than 50 distinct histological subtypes, which can occur at any age and at any location in the human body, sarcoma represents a multitude of malignancies rather than a single entity. Although as a group sarcomas exhibit unique clinical behaviors that differentiate these tumors from epithelial tumors, individual sarcoma subtypes widely differ in comparison to each other. This diversity converts sarcomas in a challenge in the sense to evaluate and analyze aspects such as the epidemiology and etiology. The design of anti-sarcoma agents with the ability to be highly active against more than one sarcoma cell line is a difficult task, taking into consideration that several sarcomas have become resistant to the current anti-tumoral drugs. Around 2000 compounds have been reported to have in principle, anti-sarcoma activity [2]. However, only few works considering approaches like QSAR techniques have been reported toward the design of anti-sarcoma compounds [3–7]. On the other hand molecular modeling techniques generally, have been more focused on aspects such discovery of new targets and mechanisms of resistance of different sarcoma cell lines as consequence of mutation in proteins [8–13].

There is no available methodology which can be able to predict in a rapid and efficient way anti-sarcoma activity. In the last 15 years, new approaches based on graph-theoretical descriptors have emerged as powerful tools for the design of bioactive agents [14–23]. The purpose of these approaches is to perform massive screenings of heterogeneous databases of compounds and to extract as much as possible structural information at different levels of chemical diversity and complexity. These approaches, supported by computer-aided drug design methods, have been developed rapidly in recent years. They have been able to predict biological activities against different target (microbial species, protein functions and others) at the same time and with a single mathematical model [24–29]. All these methodologies can be extended to the discovery and identification of new lead candidates as potential anti-sarcoma agents with a wide spectrum of action

* Corresponding authors. Fax: +351 220402659.
E-mail addresses: alejspivanovich@gmail.com (A. Speck-Planche), ncordeir@fc.up.pt (M.N.D.S. Cordeiro).

0223-5234/$ – see front matter © 2011 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2011.09.055

against different sarcoma cell lines. In this work, a unified fragment-based approach for the efficient search and design of new anti-sarcoma agents against 12 sarcoma cell lines was developed from a heterogeneous database of compounds in order to classify, design and predict anti-sarcoma agents against several sarcoma cell lines. The methodology was focused on a multi-target QSAR model by employing substructural descriptors Also, the present model was used to extract the most suitable fragments as structural alerts responsible of the anti-sarcoma activity and to design new molecular entities which were predicted by the model as possible potent and versatile anti-sarcoma agents.

2. Methods

2.1. Functional group counts

Functional group counts (FGC) are descriptors which express certain fragmental features. They are simple molecular descriptors defined as the number of specific functional groups in a molecule and they are calculated from the molecular composition and atom connectivity [30]. Many of the functional groups defined here, are those which are traditionally used in Organic Chemistry. FGC are descriptors that keep relation with the indicator variables in a Free-Wilson analysis [31].

2.2. Atom-centered fragments

Atom-centered fragments (ACF) have demonstrated to be very useful descriptors, and have been employed in some QSAR studies [32–35]. They provide important information about hydrophobic and dispersive interactions which are involved in biological processes such as transport and distribution of drugs through the membrane. Also, they give information about drug–receptor interactions [36]. ACF are simple molecular descriptors which are defined as the number of specific atom types in a molecule. They are calculated from the molecular composition and atom connectivities. Each type of atom in the molecule is described in terms of its neighboring atoms. Hydrogen and halogen atoms are classified by the hybridization and oxidation states of the carbon atom to which they are attached. For hydrogen atoms, heteroatoms which are attached to a carbon in α-position are further considered. Carbon atoms are classified by their hybridization state and depending on whether their neighbors are carbon or heteroatoms.

2.3. Spectral moments of the bond adjacency matrix

The approach that encloses the calculation of the spectral moments of the bond adjacency matrix, is known as TOPS-MODE (**TOP**ological **S**ubstructural **MO**lecular **DE**sign) approach and it has been applied for the description of some physicochemical properties of organic compounds [37–39], in quantitative structure–toxicity relationship (QSTR) [40–44], and also have been reported for the modeling of pharmacological activities [45–47]. The theoretical background about the spectral moments of bond adjacency matrix has been described in many papers; however, we will focus our explanation on the most important aspects. In this approach the molecular structure is encoded by mean of the edge adjacency matrix **E** (commonly called the bond adjacency matrix **B**). The **E** or **B** matrix is a square table of order m (the number of chemical bonds in the molecule). The elements of this matrix (e_{ij}) are equal to 1 if bonds i and j are adjacent (which means that i and j are incident in the same vertex or atom) and 0 otherwise. In order to encode information of heteroatoms, the TOPS-MODE approach uses $\mathbf{B}(w_{ij})$ weighted matrices instead of **B**. The weights (w_{ij}) are chemically meaningful numbers such as bond distances, bond dipoles, or mathematical expressions involving atomic weights. The weights are introduced in the main diagonal of matrix $\mathbf{B}(w_{ij})$. Then, the spectral moments of this matrix can be used as molecular fingerprints in QSAR studies for the codification of molecular structures. By mathematical definition, the term spectral moments must be understood as the sum of the elements (e_{ij}) in the natural powers of $\mathbf{B}(w_{ij})$ [48–50]. Then the spectral moment of order k (μ_k) is the sum of the main diagonal elements (e_{ii}) of matrix $\mathbf{B}(w_{ij})^k$. The spectral moments of the bond matrix are defined as:

$$\mu_k = Tr\left(E^k\right) = \sum_{i=1}^{s} (e_{ii})^k \quad (1)$$

where **Tr** means the trace of the matrix, that is the sum of the diagonal entries of the matrix and the elements $(e_{ii})^k$ are the diagonal entries of the kth power of the bond matrix. The spectral moments of the bond adjacency matrix have topological nature. The principal advantage of these descriptors is the possibility to calculate the relative contribution of any fragment to the desired activity [15,51]. That is possible because they can be expressed as linear combinations of the number of times in which a fragment appears in the molecule. Another advantage of the spectral moments of the bond adjacency matrix is the ability to explain in a reasonable way, a considerable part of spatial phenomena [52]. That is a particular characteristic of topographical descriptors.

2.4. Selection of the data set: calculation of the descriptors and development of the model

The data set was formed by 449 compounds with anti-sarcoma activity against 12 sarcoma cell lines [2]. Not all the compounds were tested against all the subtypes. We had also 214 drugs which have been reported in the Merck Index. These compounds present other activities that do not include anti-sarcoma activity against any sarcoma cell line and have been used as inactive [53]. The FGC and ACF were calculated using DRAGON version 5.3 [30]. The spectral moments of the weighted bond adjacency matrix (from order 1–15), were calculated using MODESLAB v1.5 [54]. In this case, the spectral moments were weighted by physicochemical properties such as polar surface area, molar refractivity, by van der Waals atomic radii and by two Abraham terms: the first containing information about the relationship dipolarity/polarizability and the second containing information about the hydrogen bond basicity. Linear Discriminant Analysis (LDA) was used to construct the classifier model. This has been a statistical technique in many QSAR studies, some of them including multi-target methodologies [26–28,55–60]. The most important step in this work was the organization of the spreadsheet containing the raw data used as input for the LDA, because this was not a classical LDA problem. For this reason we employed the multi-target QSAR methodology which was applied by González-Díaz and coworkers for the prediction of enzyme classes from 3D structure in *Leishmania infantum* [29]. Here, we formulated a 2 group discriminant function to classify compounds: compounds that belonged to a particular group (anti-sarcoma activity against a specific cell line) and compounds that did not belong to this group (inactive). For this, we had to construct an approach to define the groups predicted in each case. The following steps were used:

- First, the 449 compounds belonging to the group of compounds with anti-sarcoma activity were divided according to their activity against different sarcoma cell lines. Taking into consideration that there was an appreciable difference in activity between compounds tested against the same cell line, we decided to create cutoff values of anti-sarcoma activity