European Journal of Medicinal Chemistry 44 (2009) 3658–3664

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

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

Original article

Development of QSAR models for predicting hepatocarcinogenic toxicity of chemicals

Ilaria Massarelli [a], Marcello Imbriani [b], Alessio Coi [c], Marilena Saraceno [c], Niccolò Carli [d], Anna Maria Bianucci [c,*]

[a] Istituto Nazionale per la Scienza e Tecnologia dei Materiali, Via Giusti 9, 50121 Firenze, Italy
[b] Fondazione S. Maugeri, IRCCS, Via S. Maugeri, 4, 27100 Pavia, Italy
[c] Dipartimento di Scienze Farmaceutiche, Università di Pisa, Via Bonanno 6, 56126 Pisa, Italy
[d] Via Buonarroti 117/b, 55059 Viareggio (Lucca), Italy

ARTICLE INFO

Article history:
Received 25 July 2008
Received in revised form
29 January 2009
Accepted 12 February 2009
Available online 20 February 2009

Keywords:
Carcinogenic potency database
WEKA
Quantitative structure–activity relationship
Hepatocarcinogenic
Sphere-exclusion

ABSTRACT

A dataset comprising 55 chemicals with hepatocarcinogenic potency indices was collected from the Carcinogenic Potency Database with the aim of developing QSAR models enabling prediction of the above unwanted property for New Chemical Entities. The dataset was rationally split into training and test sets by means of a sphere-exclusion type algorithm. Among the many algorithms explored to search regression models, only a Support Vector Machine (SVM) method led to a QSAR model, which was proved to pass rigorous validation criteria, in accordance with the OECD guidelines. The proposed model is capable to explain the hepatocarcinogenic toxicity and could be exploited for predicting this property for chemicals at the early stage of their development, so optimizing resources and reducing animal testing.

1. Introduction

One of the most costly problem, when working in the early steps of discovery of new potential drugs, is related to the failure of candidates due to poor absorption, distribution, metabolism, elimination or toxicity (ADMET) properties. Recent studies attribute to ADMET problems over 60% of failures of drug candidates in development. The prediction of ADMET properties of a compound still represents a big challenge nowadays. In this perspective, several research groups working on Quantitative Structure–Activity relationship (QSAR) are emerging for the development of robust models capable to predict ahead of time these properties in order to prioritize compounds with greatest chance of success during drug discovery. Furthermore, these kind of approaches are also regarded with great interest as possible alternative methods to animal testing by a number of institutions, committees and centres of excellence like the European Centre for the Validation of Alternative Methods (ECVAM) [1]. These organizations concentrate their efforts on development and validation of alternative test methods that *refine*, *reduce* or *replace* animal usage, a principle that is comprised in the 'Three Rs' concept proposed by Russell and Burch [2].

The Organization for Economic Co-operation and Development (OECD) Group on (Q)SARs recently published (February 2007) a 'Guidance Document on the Validation of (Q)SAR Models' with the aim of providing guidance on how specific (Q)SAR models can be evaluated with respect to the OECD principles [3].

The guidelines described by OECD were considered during the development of the QSAR models described in this work. The selected end-point considered in this study is a well defined carcinogenicity index reported in the Carcinogenic Potency Database (CPDB) [4].

Nowadays, a lot of work has been done, in the field of mutagenicity and carcinogenicity, about predictions of these malignant toxicities. Unfortunately such an effort lead to a limited success. That is probably due to the fact that these end-points are very hard to be defined. Moreover the great variety of carcinogenicity mechanisms contributes to increase prediction failures. Many carcinogens are mutagenic, as they form covalent bonds with DNA, hence mutagens can be predicted by identifying electrophilic functional groups. In some cases metabolic transformations, that usually act by detoxifying the exogenous molecules, active the chemicals and produce a potential carcinogen (e.g. the N-oxidation of aromatic

[*] Corresponding author. Tel.: +390502219564; fax: +390502219605.
E-mail address: bianucci@dcci.unipi.it (Anna Maria Bianucci).

0223-5234/$ – see front matter © 2009 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2009.02.014

amines). Other carcinogens act as promoters, stimulating cell proliferation. Promoters are much more difficult to predict, as they have heterogeneous structures with diverse activities, and many are species specific.

Several QSAR works were performed in recent years with the aim of explaining carcinogenesis activity, shown both by focused chemical classes of compounds and by noncongeneric chemicals [5–11].

In this work, the homogeneity of the dataset was maximized by choosing a unique type of carcinogenic toxicity. In particular, only toxicity data referring to chemicals, which show hepatocarcinogenic potential, were chosen. Furthermore, only data referring to a unique specie (mouse) and sex (female) were taken into account during the selection of the initial dataset. Such kind of studies may be exploited to reduce animal testing in the field of carcinogenicity, by generating hazard data useful for, at least, preliminary risk assessment from exposure to chemicals and could be used within a battery of models for the prediction of this type of toxicity.

One of the main issues, to be faced when developing QSAR models, is represented by the validation step. In view of submitting the model obtained to a rigorous validation check, the whole available dataset of known molecules was split into training (TR) and test (TS) sets, according to a protocol which ensured optimal sampling both in the domain of molecular structure and molecular properties. Molecules are properly represented as points in a multidimensional space defined by molecular descriptors. Points which represent both TR and TS set molecules have to be evenly distributed within the whole descriptor space, and each point of the TS set has to be close to at least one point of the TR set. This approach ensures that the similarity principle is applied when predicting the properties of the TS set.

In this work, the rational splitting of the whole dataset into TR/TS set pairs was obtained by using a sphere-exclusion type algorithm, optimized in our lab and described elsewhere [12]. It ranks similarities among molecules before proceeding to the selection step. Similarities are calculated in terms of Euclidean distances computed on the basis of the molecular descriptors that will be subsequently exploited for model development. Molecular descriptors were computed by the CODESSA program [13], as described later in more detail.

It may be worth to point out here that the rational splitting of the whole dataset not only ensures that the TS exploited for external validation of the model contains molecules that fall into the chemical space defined by the TR set, but also defines the applicability domain of the model developed on the TR itself. As the final step, the WEKA program [14] (Waikato Environment for Knowledge Analysis) was used in order to develop the QSAR models, searched by means of many different regression algorithms. The best obtained models were then submitted to rigorous validation analysis based on several statistical parameters described in detail later on. Among them, only one model turned out to possess a very high predictive power.

2. Theoretical aspects

2.1. Rational TR/TS sets splitting

Rational splitting of the available dataset into training and test Set (TR/TS) pairs is required in order to obtain QSAR models endowed with high predictive power. Such a splitting should be performed so that points representing both TR and TS sets are properly distributed within the whole descriptor space defined by the entire dataset. The descriptor space may be defined as the multidimensional space where each one of the coordinate axes is associated to a molecular descriptor. Each molecule in the initial dataset is represented as a point in such a space. In this frame each point, representing a molecule of the TS set, should be close to at least one point of the TR set. This approach ensures that the similarity principle can be employed for the activity prediction of the TS set.

An optimized implementation of a sphere-exclusion type algorithm [15–18] previously obtained in our lab [12] was used in this work for rationally splitting the whole dataset into different TR/TS set pairs. A number of molecular descriptors computed for each molecule by using the CODESSA program [13] are handled by the algorithm which normalized them and subsequently calculated the Euclidean distances between all pairs of the dataset molecules, in the multi-dimensional descriptor space. Descriptors were normalized according to the following formula:

$$X_{ij}^{\mathrm{n}} = \frac{X_{ij} - X_{j,\min}}{X_{j,\max} - X_{j,\min}}$$

where X_{ij} and X_{ij}^{n} are the non-normalized and normalized j-th ($j = 1,\ldots, K$) descriptor values for compound i ($i = 1,\ldots, N$), correspondingly, and $X_{j,\min}$ and $X_{j,\max}$ are the minimum and maximum values for j-th descriptor. Thus, for descriptors, $\min X_{ij}^{\mathrm{n}} = 0$ and, $\max X_{ij}^{\mathrm{n}} = 1$.

By using different similarity thresholds, it is possible to select several TR/TS set pairs which are subsequently exploited for developing QSAR models.

It may be worth to point out here that each TR set defines a specific applicability domain (AD) where the model (developed on it) is expected to possess high predictive power. The criterion, underlying the rational molecule selection described above for TR/TS set splitting, must also be exploited in order to check if new chemical entities (NCE) or drugs, that are to be subjected to the QSAR model for property predictions, are comprised in the AD defined by the model itself. Only in this case property predictions for new molecules will be reliable.

2.2. Machine learning

A collection of machine learning algorithms for data mining tasks contained within the WEKA program package [14] was used for the selection of an optimal subset among all the calculated molecular descriptors and, subsequently, for the search of the best-performing algorithm during QSAR modeling.

As what concerns the group of descriptors to be used, it is worth to recall here that is should have a limited size, especially if multiple linear regression (MLR) is used. In this case, the ratio between the number of descriptor exploited and the available known molecules should be about 1:5. MLR calculates QSAR equations by performing standard multivariable regression calculations with multiple variables in a single equation. When using MLR, it is assumed that the variables belong to an orthogonal set, which is difficult to achieve in practice; nevertheless a poor correlation between variables is the condition ensuring the achievement of powerful predictive models. In this perspective, the number of independent variables initially considered should not be higher than one-fifth the number of known compounds in the training sets [19]. A higher ratio often leads to over-correlated equations, that in turn gives rise to poorly reliable predictions.

2.3. Statistical analysis and model validation

The criteria used for validating the obtained models rely on several statistical parameters that have been proved to ensure rigorous model validation. A detailed description is reported and discussed elsewhere [20]. Here it may be worth to only recall that conditions which have to be satisfied for the TR set are: $R^2 > 0.6$,