Modeling Binary Fingerprint Descriptors With the Superposing Significant Interaction Rules (SSIR) Method

Emili Besalú, University of Girona, Girona, Spain
https://orcid.org/0000-0003-0093-5714

ABSTRACT

Recently, the superposing significant interaction rules (SSIR) method has been applied in several fields of QSPR to model and establish molecular rankings that correlate dichotomous properties. The origin of the method is in the field of combinatorial chemistry, but it has been shown that the procedure is fast, versatile, and that it can be applied in many other fields. In particular, an example is phospholipidosis modeling taking, as primary descriptors, the binary fingerprints of the molecules. This is the first time SSIR is used to treat this kind of descriptors. The performance achieved is similar to other results found in the literature and, in particular, to the results obtained by authors who considered the same molecular set and descriptors. One of the main advantages of SSIR is that the method acts as an automated variable selector. This allows it to be used almost immediately without prior selection of variables.

KEYWORDS

AU-ROC Curve, Binary Fingerprints, Consensus, Phospholipidosis, Ranking, Rules, SSIR Method, Votes

1. INTRODUCTION

This paper constitutes a brief demonstration of how the Superposition of Significant Interaction Rules (SSIR) method works. In this perspective article, it is not intended to review the details and specifications of the SSIR procedure, as they are explained in several places (Besalú, 2016; Besalú et al., 2016; Besalú et al., 2017; Besalú et al., 2018). Specifically, the objective is to promote the method by showing a simple application example and its performance when handling binary fingerprints. It is the first time the method is used to rank compounds being described by binary parameters.

The molecular descriptors chosen are treated in order to model Phospholipidosis (PLD), a disorder that can be induced by several drugs. PLD is characterized by the accumulation of the inducing drug and phospholipids in the lysosomes of the affected tissues. Pharmaceutical companies need to conduct regular screenings on their drug candidates in order to avoid this side effect (Goracci et al., 2013). Several experimental and in silico methods have been described in order to predict the potential capacity of some drugs to be inducers or non-inducers (Ploemen et al. 2004; Tomizawa et al. 2006;
Pelletier et al. 2007; Kruhlak et al. 2008; Hanumegowda et al. 2010; Lowe et al 2010; Fisher et al. 2012; Lowe et al 2012; Sun et al. 2012; Orogo et al. 2012). The work of the scientific community is still in progress (Przybylak et al., 2014) regarding the definition of good predictors and even the correct classification of drugs as potential inducers. As described in the literature, it is difficult to build quality models mainly due to the fact that the phospholipidosis induction mechanism is not well known. In addition, a single curated reference molecular set is not available. It may be possible that the set proposed by the Goracci’s team is the first successful attempt to obtain it. One of the goals of this perspective article is to show how SSIR performs in this field and how it can be used to quickly rank compounds. The rankings can be used alone or can be used as preliminary molecular filters.

2. MATERIAL AND METHODS

The SSIR method is a variable selector based on the hypergeometric experiment (Mendenhall & Sincich, 1995). Briefly, given an urn containing red and green marbles, this probabilistic experiment consists of randomly selecting some of those marbles and, subsequently, assessing the probability associated with the distribution of red and green marbles which has been picked up. Specifically, the urn originally contains \( a \) marbles (\( b \) of them are green and the rest are red) and \( c \) are randomly selected. After the selection, one realizes that \( d \) of the marbles extracted are green. The probability for the described event follows the hypergeometric probability distribution:

\[
P(d; c; b, a) = \frac{\binom{b}{d}\binom{a-b}{c-d}}{\binom{a}{c}} \quad \text{with} \quad d \leq c \leq a \quad \text{and} \quad d \leq b \leq a
\]  

where the minimum allowed value for \( d \) is \( \max(0,c+b-a) \), and the maximum is \( \min(b,c) \).

The SSIR method works similarly according to the following analogies: the urn is a molecular database, the marbles are molecules and the condition of being a ‘green’ or a ‘red’ marble is equivalent to being a ‘interest’ molecule or ‘of no interest’ (or vice versa, since this classification is arbitrary). Note that the procedure applies to dichotomized molecular sets. When dealing with continuous variables, the dichotomization is defined by the user establishing a cut-off value. The process of random extraction of balls is equivalent to selecting \( a \) priori some descriptors (one or several, as it will be explained below) and, at the same time, specifying which are the levels or range of values each descriptor must have. This tandem of descriptor/s and respective levels constitutes a rule. So, a rule defines a molecular condition: a molecule will conform (fulfill) with the rule or not. All molecules conforming with the rule are (virtually) being ‘extracted’ (i.e., the rule is the extractor agent). Then, within the set of ‘extracted’ or selected molecules some will be ‘green’ and some will be ‘red’. Ultimately, the hypergeometrical Formula (1) provides the probability of obtaining the final proportion of collected molecules of each type. The ‘difficulty’ inherent to the extraction of at least a minimum number of molecules of interest is measured by a \( p \)-value equivalent to the following addition:

\[
p(d+; c; b, a) = p\left(d : \min(b,c)\right) = \sum_{i=d}^{\min(b,c)} P\left(i; c; b, a\right) = 1 - \sum_{i=\max(0,c+b-a)}^{d-1} P\left(i; c; b, a\right)
\]

where the notations \( d+ \) or, equivalently, \( d:\min(b,c) \) stand for the event that involves the extraction of \( d \) or more molecules of interest.
Estimating the Adsorption Efficiency of Sugar-Based Surfactants From QSPR Models
www.igi-global.com/article/estimating-the-adsorption-efficiency-of-sugar-based-surfactants-from-qspr-models/223214?camid=4v1a

Applicability Domain for QSAR Models: Where Theory Meets Reality
www.igi-global.com/article/applicability-domain-for-qsar-models/144689?camid=4v1a