Regulatory Toxicology and Pharmacology 71 (2015) 388–397

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

An evaluation of in-house and off-the-shelf *in silico* models: Implications on guidance for mutagenicity assessment

Robert Jolly [a,*], Kausar Begam Riaz Ahmed [b], Craig Zwickl [c], Ian Watson [d], Vijay Gombar [e]

[a] Health/Safety/Environmental, Lilly Research Laboratories, Indianapolis, IN, United States
[b] MD Anderson Cancer Center, Austin, TX, United States
[c] Non-Clinical Safety Assessment, Lilly Research Laboratories, Indianapolis, IN, United States
[d] Advanced Analytics, Lilly Research Laboratories, Indianapolis, IN, United States
[e] Drug Disposition, Lilly Research Laboratories, Indianapolis, IN, United States

ARTICLE INFO

Article history:
Received 12 November 2014
Available online 3 February 2015

Keywords:
Genotoxicity
Risk assessment
In silico models
Leadscope
Derek Nexus
Support Vector Machine
Consensus analysis
Mutagenicity prediction
Ames test
Toxicity prediction

ABSTRACT

The evaluation of impurities for genotoxicity using *in silico* models are commonplace and have become accepted by regulatory agencies. Recently, the ICH M7 Step 4 guidance was published and requires two complementary models for genotoxicity assessments. Over the last ten years, many companies have developed their own internal genotoxicity models built using both public and in-house chemical structures and bacterial mutagenicity data. However, the proprietary nature of internal structures prevents sharing of data and the full OECD compliance of such models. This analysis investigated whether using in-house internal compounds for training models is needed and substantially impacts the results of *in silico* genotoxicity assessments, or whether using commercial-off-the-shelf (COTS) packages such as Derek Nexus or Leadscope provide adequate performance. We demonstrated that supplementation of COTS packages with a Support Vector Machine (SVM) QSAR model trained on combined in-house and public data does, in fact, improve coverage and accuracy, and reduces the number of compounds needing experimental assessment, i.e., the liability load. This result indicates that there is added value in models trained on both internal and public structures and incorporating such models as part of a consensus approach improves the overall evaluation. Lastly, we optimized an *in silico* consensus decision-making approach utilizing two COTS models and an internal (SVM) model to minimize false negatives.

1. Introduction

Evaluation of the genetic toxicity of impurities introduced during chemical synthesis of pharmaceuticals has become an essential part of risk assessment intended to ensure product safety (Jacobson-Kram and McGovern, 2007; McGovern and Jacobson-Kram, 2006; Teasdale, 2011). Mutagenic potential is routinely assessed in the laboratory using the Ames assay comprised of *in vitro* bacterial systems with or without metabolic activation (McCann and Ames, 1976; Mortelmans and Zeiger, 2000). However, given the sheer number of starting materials, intermediates, impurities, and degradants involved in the process of synthesizing new drugs, it is impractical to scale-up, purify, and test each chemical entity individually in the laboratory setting. This fact underlies a compelling need to predict mutagenic potential using *in silico* models, such as quantitative structure–activity relationships (QSAR), based solely on chemical structure, which can be employed during the various stages of drug discovery and development (Kruhlak et al., 2007). The use of models is especially beneficial given the high percentage of new drug discovery efforts that never come to fruition as, in those cases, laboratory testing becomes an expensive and wasteful proposition.

Approaches employed for assessing and controlling genotoxic impurities (GTIs) in drug development have been published (Dobo et al., 2012; Dow et al., 2013) and have been shown to be highly consistent and effective. The use of *in silico* tools for mutagenicity assessment has become commonplace within this paradigm and has found support in regulatory guidances (CHMP, 2006; FDA, 2008). Recently, the ICH M7 Step 4 guidance also favors

Abbreviations: DX, Derek Nexus; LS, Leadscope; SVM, Support Vector Machine; FN, false negatives; TN, true negatives; TP, true positives; NPV, negative predictive value; PPV, positive predictive value; GTI, genotoxic impurity; SAR, structure–activity relationship; IH, in house.

* Corresponding author at: Lilly Research Laboratories, Lilly Corporate Center, Indianapolis, IN 46285, United States.
E-mail address: Jollyra@lilly.com (R. Jolly).

http://dx.doi.org/10.1016/j.yrtph.2015.01.010
0273-2300/

use of *in silico* approaches for mutagenicity prediction, stating, "Two (Q)SAR prediction methodologies that complement each other should be applied," and that, "One methodology should be expert rule-based and the second methodology should be statistical-based" (ICH Harmonised Tripartite Guideline Step 4, 2014). Moreover, the guidance indicates that if warranted expert judgment should be applied to any evaluation (ICH Harmonised Tripartite Guideline Step 4, 2014). ICH guidance will undoubtedly direct the pharmaceutical and other industries to begin employing at least two *in silico* models for assessing the potential mutagenicity of GTIs.

Eli Lilly and Company has routinely used both a commercially available model, such as Derek Nexus (DX 3.01; Lhasa LTD, 2014 described in Greene et al., 1999), and an internally-developed QSAR model to assess potentially genotoxic impurities. In addition to including publicly available compounds in the QSAR training set, the internal model benefits by including molecules from our in-house compound library that have previously been tested in the Ames assay. The overall formal evaluation also includes a review by an expert. A retrospective analysis of our current process for GTI risk assessment using actual laboratory test data (180 compounds) has shown 100% negative predictive value (NPV – compounds predicted negative and subsequently confirmed as negative), with no false negatives (Dobo et al., 2012).

In the light of the ICH M7 Step 4 guidance, we were interested in evaluating the performance of two commercial off-the-shelf (COTS) *in silico* prediction tools, namely the rule-based Derek Nexus 3.0 expert system from Lhasa Limited (DX) expert system and statistical-based Leadscope's Ames mutagenicity (LS) QSAR model in its Genotoxicity Suite v 3.1.1–10 (Leadscope, 2014). These COTS products, as supplied by the vendors, were evaluated to determine (1) whether these individually and/or in combination provide sufficient predictive accuracy to adequately evaluate mutagenicity risk potential of our in-house compounds, and (2) whether model predictivity can be enhanced by including proprietary structures in their training sets.

To evaluate the performance of COTS products, we compiled and curated several datasets from public data sources and in-house Ames test results. We developed a Support Vector Machine (SVM) mutagenicity predictor using the same training set as used by the LS model to compare the two modeling methods (LS & SVM). We also developed several SVM models with training sets comprising both the publicly available and in-house Ames data to determine whether significantly better performance would result by adding proprietary chemical space to training sets. Finally, we were interested in determining the optimal configuration of models for achieving a balance between predictive accuracy (minimization of false negative predictions in particular) and the resources required to conduct non-essential Ames studies that would not compromise the integrity of a rigorous risk assessment approach.

It should be noted that the present evaluation focuses purely on the performance of the *in silico* models in the absence of the expert judgment component of our complete GTI process. The driver for this analysis was to enhance and optimize the *in silico* component of the process as much as possible. The reported evaluation is intended to aid discussions on the application of, and expectations for, the use of *in silico* approaches as part of the safety–risk profiling of pharmaceuticals and residuals.

2. Data sets and models

Table 1 describes the data sets used to train or test the performance characteristics of the models. The table also describes the COTS models and SVM models developed in-house.

2.1. Public data sets

LS_3970 is the dataset used to train the LS model. In addition to this dataset, we compiled and curated an extended dataset of Ames results from publicly available sources such as the Vitic Nexus database (Lhasa Ltd., 2014), and several literature sources (Hansen et al., 2009; Kazius et al., 2005; Feng et al., 2003), the Carcinogenic Potency Database (Swirsky, 2011), and the Physician's Desk Reference (PDR). The combined set was curated by excluding inorganic and organometallic compounds, retaining the free acid/base form of any compound tested as a salt, and excluding duplicate structures. The final set contains 8541 unique structures with unambiguous Ames assay results and is designated Public_8541 (Table 1). It was used both as a test set to characterize the overall performance of models on public compounds and as a training set for several SVM models, either alone or after supplementation with in-house structures. There was overlap in the public datasets and in the Ames results. In the Public_8541 set, 3174 compounds were also present in LS_3970. There were 164 compounds with conflicting calls. We retained the calls in LS_3970 as the "true" calls. When there was a conflict in the public vs. in-house data, we took the in-house call. When there was a conflict between in-house and LS_3970, we used the LS_3970 call, and when there was a conflict between Public_8541 and Vitic_3863 that compound was not included in the analysis.

The Vitic Nexus dataset contains 3863 publicly available compounds and is a subset of structures used by Lhasa, Ltd., to develop DX alerts for its Knowledge Base. This set is designated Vitic_3863 (Table 1) and, like the Public_8541 set, was used as a test set for evaluating model performance. The Vitic_3863 set also overlapped with Public_8541 set and again, compounds with conflicting data were not included in the analysis. The number of mismatched calls was <5% relative to the number of compounds in the Public_8541 set. The Vitic set was not a focus of the paper but rather was included to benchmark how DX performed on compounds in the Vitic database.

2.2. In-house data sets

Two thousand thirty-six unique compounds have been assayed in the Ames test at Lilly since 1979. This set was divided into a subset of 1605 training compounds (In-House_1605, Table 1), used alone or to supplement public data sets, and a subset of 438 test compounds that were most recently (since 2011) tested in the

Table 1a
Description of datasets used.

Dataset	Type	Number of compounds	Ratio of non-mutagens to mutagens	Data source
LS_3970	Training	3970	1.32	Leadscope
Public_8541	Training	8541	1.05	Our compilation of data from public sources
In-House_438	Test	438	3.47	Recent in-house compounds with Ames data
In-House_1605	Training	1605	5.38	In-house compounds with Ames data since 1979 which does not include In-House_438
Vitic_3863	Test	3863	1.04	5-Strain data from Lhasa Ltd.