Predicting Skin Sensitizers with Confidence - Using Conformal Prediction to Determine Applicability Domain of GARD®

Henrik Johansson, 1,2,3 Andy Forreyd, 1,2 Ulf Norinder, 1,2,3 Tim Lindberg, 1,2,4 Malin Lindstedt, 1,2,3 BengtGen AB, Lund, Sweden, 2Swetox, Karolinska Institutet, Södertälje, Sweden, 3Department of Computer and Systems Sciences, Stockholm, Sweden, 4Department of Immunotechnology Lund University, Lund Sweden.

OBJECTIVE

GARD® – Genomic Allergen Rapid Detection – is a state of the art non-animal based technology platform for classification of skin sensitizing chemicals. The assay has proven to be reliable and highly accurate for identification of skin sensitizing chemicals, and consistently reports predictive performances > 90% across external test sets. The aim of the current project is to complement assessments of average model performance with an estimate of uncertainty involved in each individual prediction, thus allowing for classification of skin sensitizers with confidence.

THE GARD ASSAY – BACKGROUND

Genomic Allergen Rapid Detection classifies chemicals as either skin sensitizers or non-sensitizers using a Support Vector Machine (SVM) model trained on the readout from a genomic biomarker signature of 200 genes (GARD® prediction signature, or GPS) measured in an in vitro model of dendritic cells. The assay was recently subjected to a formal ECVM validation procedure (OECD TGP 4.106) performed in accordance with OECD guidelines. The study involved classification of 28 tests substances across three independent laboratories and demonstrated an excellent reproducibility within (82-89%) and between laboratories (92%), as well as an outstanding predictive performance: accuracy - 94%, sensitivity – 93%, and specificity 96%. The assay procedure of the GARD® assay is illustrated in figure 1.

CONFORMAL PREDICTION

Predictive accuracy based on classification of external test sets is often the only performance metrics reported for in vitro assays. This metrics does not inform on the uncertainty involved in prediction of individual test substances which may occur when a model is forced to classify a test substance falling outside the applicability domain of the assay. Conformal Prediction (CP) is a mathematical framework that provides a measure of uncertainty associated with each classification, as well as deliver a warning if a specific test sample falls outside the applicability domain of the assay.

METHODS

In CP, applicability domain is defined based on similarity to training samples. In GARD®, which uses an SVM model for classification, the distance to the separating hyperplane was used as a measure of conformity. To estimate how similar a test compound was to training samples, the conformity value of the test sample was compared to conformity values for samples used for model training, by calculating a CP P-value (fraction of samples in the training set with a lower conformity value compared to the test sample) for each class label. To decide if a test sample was similar enough to training samples, the P-value for each class label was compared to a user defined significance level (α), and assigned to (if p > α) either one of the two classes (single), to both classes (model was unable to distinguish between classes), or to the empty domain (model was unable to assign class label, since compund was too dissimilar to samples in training set, and thus by definition outside applicability domain). The overall strategy of CP is illustrated in figure 3 and figure 4.

RESULTS

An Internal validation procedure was initially performed on samples in the GARD® training set (n=38) using the strategy described in Fig. 3A. Results from this exercise is summarized in Fig. 4A. Conformal prediction by definition allows the user to determine a reasonable and acceptable significance level to guarantee a maximum error rate in predictions. The significance level was set to 15%, i.e. the model was allowed to make a maximum of 15% errors. Performance of the conformal predictor was measured by validity and efficiency. A model was valid if the number of prediction errors did not exceed the significance level, while efficiency corresponded to the percentage of single class predictions. Internal validation of the training data resulted in a valid and highly efficient model (92% single classifications, 1 empty, 2 both), indicating that the ambitious significance level was at a reasonable level for the GARD® assay. Following internal validation, samples in a large external test set (n=70) was classified within the CP framework as described in Fig. 3B, which resulted in generation of a valid and highly efficient model (99% single classifications, 0 empty, 1 both) (Fig. 4B). Additional data on model performance is illustrated in Table 1.

CONCLUSION

This study addressed the important aspect of defining applicability domains of in vitro assays, which is often neglected by test developers. To address this issue, Conformal Prediction was implemented into GARD® settings. An external test set containing all chemical reactivity domains was classified and results showed that the multivariate biomarker signature used in GARD® is both applicable, as well as highly accurate, across a large chemical space, including pre- and pro- hapten, as demonstrated by few samples assigned into empty and both domains. For regulatory bodies, as well as for the chemical industry, this results facilitate a decision making process, and allow for classification of skin sensitizers with confidence.

Acknowledgement

The research at Lund University was supported by grants from the Swedish Foundation for Strategic Research (IB13-0118) and the Faculty of Engineering (LTH). The research at Swetox was supported by Knut & Alice Wallenberg Foundation (20120523) and Swedish Research Council FORMAS (2016-02031).

Additional reading


Figure 1. GARD® – Genomic Allergen Rapid Detection. Chemically induced changes in transcriptional levels of genes expressed to test chemicals are compared to predictive genomic biomarker signatures using Support Vector Machine (SVM) algorithm based on supervised machine learning. Unknown samples are classified as either skin sensitizers or non-sensitizers based on output from the machine learning algorithm.

Data [empty] (#) [both] (#) Efficiency (%) Validity (%) Accuracy (%) Balanced Accuracy (%)
NS S NS S
Training (n=38) 1 2 92 95 83 95 87 91
Test (n=70) 0 1 99 94 83 94 83 88

Table 1. Additional data on model performance.

Figure 2. Illustration of the CP concept using simulated data. An SVM model was trained on samples in training set and applied to assign conformity values to samples in a calibration set (black and open circles, resulting values summarized in the table), and samples in external test set (grey circles). P-values were generated for each test sample by determining the fraction of samples in the calibration set with < conformity value compared to the test sample. Resulting p-values (P) for compound 1 (CV=35%), compound 2 (CV=35%), and compound 3 (CV=100%) were calculated to 0.61, 0.61 and 0.61 for sensitization class, and 0.73, 0.73, and 0.73 for non-sensitization class, respectively. Thus, at significance level 15% 0.15, compounds in external test set were classified as non-sensitizer (1), sensitizer (2) and non-sensitizer (3).