VEGA-QSAR: AI inside a platform for predictive toxicology

Emilio Benfenati[1], Alberto Manganaro[1] and Giuseppina Gini[2]

[1]IRCCS- Istituto di Ricerche Farmacologiche Mario Negri, Milano, Italy
{benfenati, manganaro}@marionegri.it

[2]DEIB, Politecnico di Milano, Italy
giuseppina.gini@polimi.it

Abstract. Computer simulation and predictive models are widely used in engineering, much less considered in life sciences. We present an initiative aimed to establish a dialogue within the community of scientists, regulators, industry representatives, offering a platform which combines the predictive capability of computer models, with some explanation tools, which may be convincing and helpful for human users to derive a conclusion. The resulting system covers a large set of toxicological endpoints.

1 Introduction

Predictive toxicology is using models to predict biological endpoints, in particular toxicity, without making real experiments. The concept of "Structure-Activity Relationship" (SAR) is that the biological activity of a chemical can be related to its molecular structure. When quantified, this relationship is known as "QSAR". A QSAR model makes use of existing experimental toxicity data for a series of chemicals to build a model that relates experimentally observed toxicity with molecular descriptors in order to predict the toxicity of further chemicals. The term predictive toxicology [1] has been introduced by the AI community to indicate this approach to toxicology.

A series of regulations require producing information about the safety of the chemical substances, such as the European legislation REACH. This regulation states that for each chemical circulating in Europe a complete dossier on physico-chemical, environmental and toxicological properties has to be compiled. In order to prevent an over-usage of animal testing, REACH foresees the use of alternative methods, including predictive programs.

Life sciences are heavily impacted by the development of methods for data collection and analysis; they are moving from an analytical approach to a modelling approach. An important step in this direction is played by the changing mind from purely statistics usage of data to the data mining and machine learning view of the

recent years. Good models should (1) explain patterns in data; (2) correctly predict the results of new experiments or observations; and (3) be consistent with other ideas (models, beliefs). Of course the (3) requirement is the most critical one.

Today some QSAR models have proved to offer a valuable alternative to the classical *in vivo* methods [2]. Nevertheless, most of the QSAR models are not trusted by their targeted users, for several reasons [3], including the misunderstanding of the technology. We decided to cope with those issues through an open platform dedicated to the stakeholders potentially interested in using QSAR models.

This paper presents the developed platform. The user needs, captured during several workshops, interviews and exercises carried on through four recent European projects (DEMETRA, CAESAR, ORCHESTRA and ANTARES) have strongly guided the platform development. According to the user's requirement to keep as confidential the chemical structures, our solution is implemented both as a web-based application and as down-loadable software.

2 Using the VEGA-QSAR platform

Several institutes contributed to the development of the platform, called VEGA-QSAR, including regulators and public bodies in Europe and USA. VEGA freely offers tens of models for properties such as persistence, logP, bioconcentration factor (BCF), carcinogenicity, mutagenicity, skin sensitization.

The initial nucleus of VEGA models derives from the CAESAR models[1]. Other models have been added to simulate the models developed by the partners; this is the case of models developed by EPA (US Environmental Protection Agency) and ISS (Istituto Superiore di Sanità) for instance. All the models have been published in scientific literature before incorporating into VEGA. Moreover all the models have been successfully benchmarked against the few commercially available systems.

The steps of the workflow are clearly indicated in the GUI: insert the list of the molecules identifiers, choose where to send the prediction output, ask prediction, and get results. The input can be given in different standard formats used in the chemical domain, including SMILES and SDF files [4]. To avoid the well-known problems about the non-unique representations VEGA transforms all the chemical structure into a unified internal string format.

Figure 1 show, as an example, the output screen of the BCF model [5] with the prediction and the most similar compounds with their experimental and predicted values. BCF is a dose value, however for regulation classes are assigned according to thresholds. Since the uncertainty of the prediction can be calculated, it is graphically shown for each molecule as a worst-case analysis, as in Figure 2.

[1] http://journal.chemistrycentral.com/supplements/4/S1

Furthermore, the model provides other pieces of information useful to the evaluator, such as a plot showing the experimental values of the training set (Figure 3), to check for possible unusual behaviour of the target compound.

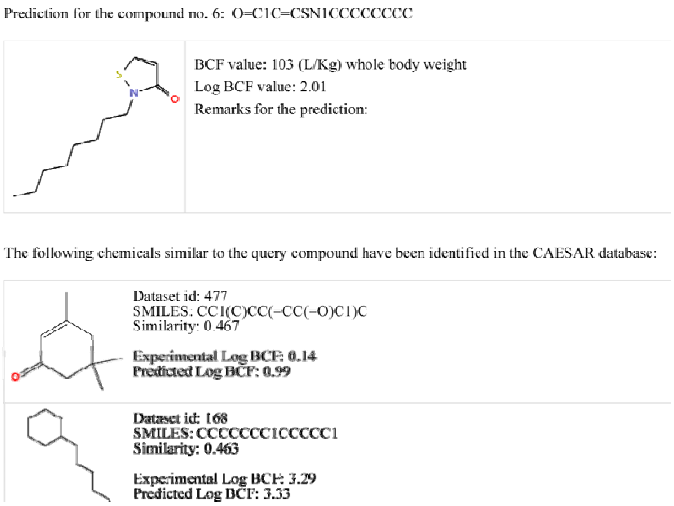

Figure 1 – Prediction of the bioconcentration factor (logBCF) for the compound and the most similar structures available in the dataset, with experimental and predicted values.

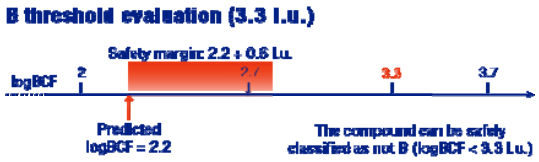

Figure 2 – The BCF model suggests the classification in the classes defined in REACH**.**

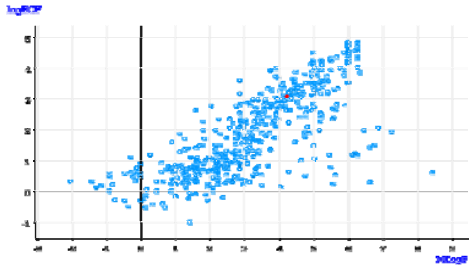

Figure 3 – The predicted logBCF value (red dot) inside the experimental values of the dataset.

The overall reliability of the prediction is measured by combining statistical values, elements of case based reasoning, and possibly presence of active substructures; the possible reasons of concern are underlined. All those considerations are weighted and summed up in an index (in 0 – 1) that is called Applicability Domain Index (ADI). Moreover the user can apply different models to the same endpoint, and VEGA suggests the best integration, as illustrated in Section 5.

3. AI inside VEGA

VEGA is mainly the result of the data driven and knowledge discovery approaches. The models implemented arise from different methods. At least three families are represented: (1) rule based expert systems, usually codified as the presence of chemical substructures called structural alerts (SA), defined by experts, as in Toxtree [7]; (2) data miners that extracts relevant fragments from the analysis of their correlation with the endpoint, as in SARpy [6]; (3) regression models that use molecular descriptors and non linear methods - ANN, SVM - as the mentioned BCF model; (4) ensemble methods as random forests as in Developmental toxicity; (5) hybrid models that mix the above methods as in CAESAR-mutagenicity.

The toxicity domain poses significant challenges to the AI methods. The first is the lack of available knowledge about reasons and mechanisms of toxicity for many of the endpoints that make it impossible to apply deductive approaches. The purely symbolic approach can be used for “mechanistic interpretation”, which is a vague indication about the toxicological pathway associated to that chemical on the basis of the presence of a specific chemical subgroup. Moreover, even in case some SAs are known, their presence is only a sufficient condition for toxicity; so the absence of SA in a molecule does not guarantee its safety.

This is the reason why we have developed new methods as in SARpy [6], with the aim of discovering both new SA and ”neutral substructures” (NS) of the molecule that may reinforce the classification as safe for molecules not containing SA but instead containing some NS.

QSAR systems make more use of probabilistic AI than symbolic AI, with the consequent problems of difficulty in understanding the results. There is generally a trade-off between prediction quality and interpretation quality of a model. Interpretable models are desired to make expert decisions; however those models suffer because the generalizations necessary to get them may be flawed by lack of enough data. To avoid the risk of excess generalization, often QSARs are simple linear regression in the small population of a chemical class. Those models have no predictive value outside this small population.

In VEGA models are generally not intended to provide transparency in se, but high accuracy on new data that the model has not used in training; since transpar-

ency is needed, this is obtained adding extra visualization and explanation features to the models, as presented in the previous subsection.

What is needed is a way to predict new chemicals and to deal with real substances that generally are mixtures of quite complex molecules, as in dyes and fuel, and that can be better modelled as large SAs and NSs.

4 Chemical space analysis tool

The identification of chemical features and their role in the prediction offer important modulating information about models. The idea is to find, for each model, a set of chemical features (i.e. molecular fragments or functional groups) that can identify a chemical class, and that show a statistically relevant correlation with the reliability of the prediction. If a set of chemical features are found only in compounds that yield accurate predictions, or vice versa non-accurate, this finding can provide also a more chemical-oriented explanation of the model's behaviour.

Two sets of fragments have been considered and implemented in VEGA and freely available[2]: Functional Groups (FC) that account for 154 chemical groups, and Atom-Cantered Fragments (ACF), for 115 fragments, each one corresponding to a type of atom with different connectivity. These fragments have been studied by experts in the fields of molecular descriptors [8], and have precise chemical meaning.

The software to analyse the chemical space checks for the presence of the above mentioned FG and ACF, then reports, for each of these chemical features, the total number of matches, the number of matches in each class, and its percentage. For each model, it t is possible to check the percentages of correct predictions for those fragments.

For example, for BCF all the compounds used in the model have been labelled as: correct prediction (the possible error is [-1, 1] log units), under-estimated (the predicted value is lower of at least 1 log units with respect to the experimental value), over-estimated (the predicted value is higher of more than 1 log units). We discovered that BCF does not always provide reliable predictions when more than one halogen atom is in the molecule. Moreover multiple halogens, and in particular multiple Fluorine atoms, lead to under estimated predictions.

Considering the CAESAR mutagenicity model, we discovered that the presence of hydroxylamines in a molecule leads to unreliable predictions; in particular in case of multiple hydroxylamines. Furthermore, this chemical feature is mostly related to false negatives, so it could be used as an alert. Instead halogenated compounds mostly give false positive predictions, and this can generate a warning.

The analysis of the molecule in the chemical space of each used model gives extra information to the stakeholder.

---

[2] http://kode-solutions.net/en/software_ist_chemfeat.php

5 Integrated models

During the ANTARES project we collected large data sets for some toxicity endpoints and benchmarked the available models. Since more models are implemented in VEGA for the same endpoints, we have devised different integration strategies according to the characteristics of models.

1 - *Trust the best model*. If one strong model clearly outperforms the others, the integration suggests:

- use the best model; check for possible mechanistic interpretation;
- if no prediction from the best model, then trust the others according to their applicability domain index value.

This strategy applies to ready biodegradability, where VEGA-SARpy outperforms the others, as in the ROC space in Figure 4. However BIOWIN will provide good results for chemicals out of the applicability domain of SARpy.

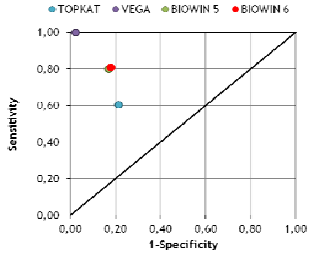

Figure 4 - ROC for molecules in test set for some ready biodegradability models

2 – *Take the best model and refine with chemical and SAs information***.** If the best model has a low accuracy, use it and check it for the chemical class and possibly for SAs.

This applies to carcinogenicity, where CAESAR model outperforms the others, as in the ROC of Figure 5, but is not very strong. The integrated strategy is.

- Take the prediction of CAESAR and check with Toxtree for possible SAs; if the results agree, accept, otherwise refine considering chemical classes

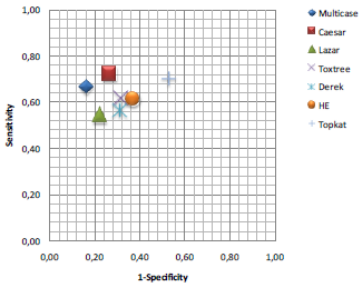

Figure 5 – ROC points for some carcinogenicity models.

3 – *Build a decision tree*. When two regression models have good performance and use different methods, it is wise to use them together.

This applies to BCF, to integrate the two VEGA models CAESAR and Meylan. The steps are:

- Compare the two values predicted by CESAR and Meylan.
- If their difference is less than 1 log unit, then exit with the highest value;
  else consider the 2 ADI indexes;
  - if none of them is greater than 07, then discharge the prediction and exit,
  else chose the predicted value with highest ADI index.

This integrated model [9], considering only the molecules in the new ADI, has R^2 =0.92, sensitivity of 85%, RMSE of 0.44 log units, with a considerable improvement from both the starting models

4 – *Make a mechanistic integration*. For good models it is possible to find a mechanistic integration, now considering the available SA-based models, in future automatically constructing them, on demand, using SARpy.

We see how this applies to mutagenicity. We run SARpy and CAESAR models. In case there is a consensus and the ADI is > 0.8, prediction is strongly supported. Otherwise, the SAs given by SARpy are to be considered. In this integration we obtain both high reliability and mechanistic interpretation.

5 – *Make statistic integration***.** In this case various models are ensembled through a gating network.

We see it again for mutagenicity that having a large dataset of more than 6000 molecules is a good candidate to statistical integration. We used a self-organizing tool of KnowledgeMiner Software[3], based on GMDH neural networks. For ensemble development the dataset was randomly half subdivided into training and testing datasets. For each compound the input vector contains the experimental class label, the prediction of the SARpy model and its applicability domain index, the predictions of the CAESAR model and its applicability domain index. The overall accuracy of the ensemble model is 88.6%, with sensitivity of 89.7%, and specificity of 87.5% that increases the classification accuracy of up to 10% relative to the individual models.

6 Final remarks and conclusions

In the development of the VEGA platform we analysed the needs of the users, and the barriers to the use of computers. We made AI approaches working in the language of the user, in order to improve understanding and acceptance of the technology [10].

[3] http://www.knowledgeminer.com

Stakeholders are continuously involved, and their feed-back used to improve the platform. Recently, an exercise [11] with tens of experts demonstrated that human experts identified reasons of possible concern evaluating the results of some QSAR models. Comparing VEGA with other platforms, the users appreciated the details and supporting information provided by VEGA.

Another important aspect is that models are inserted into VEGA-QSAR after an independent scientific evaluation, which guarantees not only their validity but also their ability to emulate the in-vivo tests results. Please keep in mind that the reproducibility of in-vivo tests is never 100%; for difficult endpoints as carcinogenicity a value less than 70% is reported, for mutagenicity 85% is reported. Predictors that have a similar accuracy have application potency.

VEGA at http://www.vega-qsar.eu contains the web application and the downloadable software. So far more than one thousand users downloaded it.

Acknowledgments

We acknowledge support of the EU Life+contracts CALEIDOS and PROSIL.

References

1. Gini G, Katritzky A. (Eds.) (1999) Predictive Toxicology of Chemicals: Experiences and Impact of Artificial Intelligence Tools. AAAI Spring Symposium on Predictive Toxicology, AAAI Press, Menlo Park, California.
2. Benfenati E., Crètien J.R., Gini G., Piclin N., Pintore M., Roncaglioni A. (2007) Validation of the models. In Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes, Elsevier, 185-200.
3. Benfenati E. et al. (2011) The acceptance of in silico models for REACH: Requirements, barriers, and perspectives. Chemistry Central Journal, Vol 5:58, 1-11.
4. Weininger D. (1988) SMILES a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chemical Information and Computer Science 28, 31–36.
5. Lombardo A., Roncaglioni A., Boriani E, Milan C., Benfenati E. (2010) Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish. Chemistry Central Journal, Jul 29; 4 Suppl 1:S1.
6. Gini G. et al. (2013) Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction. SAR and QSAR in Environmental Research, Vol 24, N 5, 365-383.
7. Toxtree (2012) software download: http://sourceforge.net/projects/toxtree/
8. Todeschini R., Consonni V.(2009) Molecular Descriptors for Chemoinformatics. Wiley-VCH.
9. Gissi A. et al (2013) Integration of QSAR models for bioconcentration suitable for REACH. Science of the Total Environment 456-457, 325–332.
10. Gini G. (2013) How Far Chemistry and Toxicology Are Computational Science?. in Amigoni and Schiaffonati (eds): Methods and experimental techniques in computer engineering. Springer, 15-33, ISBN 978-3-319-00272-9 .
11. Benfenati E. et al. (2013) Using toxicological evidence from QSAR models in practice. Altex, Vol. 30, 19-40.