Lazar Toxicity Predictions

Carcinogenicity - Mouse carcinogenicity (both sexes)

Endpoint Definition

Active if at least one target site has been reported, inactive if no positive results have been reported [Details] [Original data]

Algorithm Definition

`lazar` obtains predictions from the experimental results of compounds with similar structures (*neighbors*). For differentiated predictions chemical similarities are always determined *in respect to the endpoint under investigation* . A detailled description and formal definition of the `lazar` algorithm has been published in:

- C. Helma: Lazy Structure-Activity Relationships (lazar) for the Prediction of Rodent Carcinogenicity and Salmonella Mutagenicity, Molecular Diversity 10, 147-158 (2006) [preprint]

The present version of `lazar` uses a slightly modified definition for chemical similarity that uses a) a gaussian distribution function and b) considers the presence of fragments that cannot be evaluated for statistical reasons (i.e. because they are too infrequent in the database). The definition for chemical similarity (Equation 1) is now

$$sim(x_q, x_i, D) := \frac{\sum_{f \in F} \{e^{-(c^2/2\sigma^2)} | f \subseteq x_q \wedge f \subseteq x_i\}}{\sum_{f \in F} \{e^{-(c^2/2\sigma^2)} | f \subseteq x_q \vee f \subseteq x_i\}} \cdot \frac{n_{x_q}^{f_{frequent}}}{n_{x_q}} \cdot \frac{n_{x_i}^{f_{frequent}}}{n_{x_i}} \quad (1)$$

with
$p_f | f \subseteq x_q \wedge f \subseteq x_i$ ... significance of fragment f that occurs in x_q and x_i
$p_f | f \subseteq x_q \vee f \subseteq x_i$ ... significance of fragment f that occurs in x_q or x_i
σ ... standard deviation of the gaussian distribution (0.3)
F ... set of significant features
n_{x_q} ... number of fragments in the query structure
$n_{x_q}^{f_{frequent}}$ ... number of query structure fragments that occur frequently enough for statistical evaluation
n_{x_i} ... number of fragments in the neighbors
$n_{x_i}^{f_{frequent}}$ ... number of neighbors fragments that occur frequently enough for statistical evaluation
Minimum frequencies for statistical significance are derived from the χ^2 definition (with Yates correction) under the assumption that the fragment occurs only in asingle class.

You can donwload the source code for this `lazar` version (GNU General Public License) with git :

```
git clone http://opentox.org/git/ch/lazar-core.git
```

Applicability Domain Definition

The applicability domain (AD) of the training set is characterized by the confidence index of a prediction (high confidence index: close to the applicability domain of the training set/reliable prediction, low confidence: far from the applicability domain of the trainingset/unreliable prediction). The confidence index considers (i) the similarity and number of neighbors and (ii) contradictory examples within the neighbors. A formal definition can be found in:

- C. Helma: Lazy Structure-Activity Relationships (lazar) for the Prediction of Rodent Carcinogenicity and Salmonella Mutagenicity, Molecular Diversity 10, 147-158 (2006) [preprint]

The reliability of predictions decreases gradually with increasing distance from the applicability domain (i.e. decreasing confidence index). Figure 1 shows this dependency visually, Table 1 weights true/false predictions with their confidence and provides the best indication of the overall performance of the system.

For simplicity we provide also results for an applicability domain definition with a sharp border at a confidence index of 0.025 . These results are summarized in Table 2 , indicated by the grey area in Figure 1 and in the ROC curve in Figure 2. . Misclassifications within the applicability domain are summarized in the table of misclassifications .

The presence of substructures that are unknown to the training set (*unknown fragments*) is another factor that limits the applicability domain. As the training data cannot provide any information about unknown fragments, their relevance has to be evaluated by an expert user (as a rule of thumb large fragments are of less concern, because all shorter subfragments have been evaluated by the system). For this reason the presence/absence of unknown fragments is not considered in the formal applicability domain definition, but their presence is indicated in the table of misclassifications .

Validation Results (leave-one-out crossvalidation)

Definition and experimental comparison with external validation procedures:

- R. Benigni, T. I. Netzeva, E. Benfenati, C. Bossa andR. Franke, C. Helma, E. Hulzebos, C. Marchant, A. Richard, Y.-T. Woo, and C. Yang. The expanding role of predictive toxicology: an update on the (Q)SAR models for mutagens and carcinogens. J Environ Sci Health C Environ Carcinog Ecotoxicol Rev., 25:53-97, 2007.
- C. Helma: Lazy Structure-Activity Relationships (lazar) for the Prediction of Rodent Carcinogenicity and Salmonella Mutagenicity, Molecular Diversity 10, 147-158 (2006) [preprint]
- C. Helma and J. Kazius: Artificial Intelligence and Data Mining for Toxicity Prediction, Current Computer-Aided Drug Design 2, 1-19 (2006) [preprint]
- Presentation at Workshop on Evaluating Prediction Models in Mutagenicity and Carcinogenicity, Rome, Italy (2006) [presentation]

Table 1: Predictions weighted by confidence index

True positive predictions	tp	23.44
True negative predictions	tn	27.69
False positive predictions	fp	7.05
False negative predictions	fn	8.59
Sensitivity (true positive rate)	tp/(tp+fn)	0.73

Specificity (true negative rate)	tn/(tn+fp)	0.8
Positive predictivity	tp/(tp+fp)	0.77
Negative predictivity	tn/(tn+fn)	0.76
False positive rate	fp/(tp+fn)	0.22
False negative rate	fn/(tn+fp)	0.25
Accuracy (concordance)	(tp+tn)/(tp+fp+tn+fn)	**0.77**

Best indication of the overall performance (see Applicability Domain Definition)

Table 2: Predictions within applicability domain

True positive predictions	tp	194
True negative predictions	tn	241
False positive predictions	fp	84
False negative predictions	fn	88
Sensitivity (true positive rate)	tp/(tp+fn)	0.69
Specificity (true negative rate)	tn/(tn+fp)	0.74
Positive predictivity	tp/(tp+fp)	0.7
Negative predictivity	tn/(tn+fn)	0.73
False positive rate	fp/(tp+fn)	0.3
False negative rate	fn/(tn+fp)	0.27
Accuracy (concordance)	(tp+tn)/(tp+fp+tn+fn)	**0.72**

Predictions with a confidence > 0.025 are considered to be within the applicability domain (see Applicability Domain Definition)

Table 3: All predictions

True positive predictions	tp	261
True negative predictions	tn	329
False positive predictions	fp	135
False negative predictions	fn	138
Sensitivity (true positive rate)	tp/(tp+fn)	0.65
Specificity (true negative rate)	tn/(tn+fp)	0.71
Positive predictivity	tp/(tp+fp)	0.66
Negative predictivity	tn/(tn+fn)	0.7
False positive rate	fp/(tp+fn)	0.34
False negative rate	fn/(tn+fp)	0.3
Accuracy (concordance)	(tp+tn)/(tp+fp+tn+fn)	**0.68**

Poor indication of the overall performance. Depends predominatly on the fraction of compounds beyond the applicability domain, which are by definition poorly predictable (see Applicability Domain Definition)

Figure 1: Cumulative accuracy vs. prediction confidence

Depicts the dependency of predictive accuracy on the confidence index (i.e. the distance to the applicability domain, see Applicability Domain Definition). Fluctuations at the left hand side of the figure are statistical

artefacts (small sample sizes) and therefore irrelevant.

Figure 2: Receiver operating characteristic (ROC)

Depicts true versus false positive rates. An optimal model would reside in the top left corner, random guessing would lead to point near the diagonal line.

The table of misclassifications shows all misclassified instances within the applicability domain.

Mechanistic Interpretation

Neighbors

Neighbors are compounds that are similar in respect to mouse carcinogenicity (both sexes) . It is likely that compounds with high similarity act by similar mechanisms as the query compound. You can retrieve additional experimental data and literature citations for the neighbors and the query structure by following the "Search PubChem" links on the prediction page.

Fragments

Activating and deactivating parts of the query compound are highlighted in red and green. Fragments that are unknown (or too infrequent for statistical evaluation are marked in yellow. You can retrieve additional statistical information about the individual fragments by following the "Relevant Fragments" link. Please note that `lazar` predictions are based on neighbors and not on fragments. Fragments and their statistical significance are used for the calculation of activity specific similarities.

Built with: lazar, OpenBabel, CDK, Ruby on Rails

JME Editor courtesy of Peter Ertl, Novartis