Internal validation in OASIS models for genotoxicity

(Bacterial reverse mutation Ames with S9)

The following is a description of the process of internal validation in OASIS models for genotoxicity providing binary predictions (positive/negative).

Internal validation

Once a model is derived its predictive performance is measured by the goodness-of-fit (GOF), i.e. how well the model fits its observed data. However, it is known that this measure is optimistic and is higher than the predictive performance over any other data different from the training set. If an external (test) data is available than an *external validation* could be done and the performance over the test set could be compared to the goodness-of-fit.

If no external data is available a statistical procedure called *internal validation* could be executed in order to assess how well the model would predict external data. During this procedure the original training set is separated many times randomly into two parts – one becomes a training set and the other becomes a test set. A new model is derived for each new training set and the new model's performance over the both training and test sets is registered. The averaged value of all training set performances is compared to the averaged value of all test set performances in order to assess the amount of optimism in the goodness-of-fit (GOF optimism, or overfitting) in the original model.

Selecting the training and test samples

Different procedures exist for selecting the training and test samples. Each of them has its pros and cons, so here we will use two acknowledged methods – cross-validation and bootstrapping.

In ***k-fold cross-validation*** the original training set is partitioned into *k* equally sized subsets. Each time a single subset is used as a test set and the remaining *k-1* subsets are used as training set. In this manner the process is repeated *k* times and each data from the original training set is used once as a test data and *k-1* times as a training data. The advantage of this method is that any data is used for both training and validation and each data is used exactly once as a test data. Commonly the 10-fold cross-validation is used (90% training data, 10% test data), but we will use also 4-fold cross validation (75% training data, 25% test data).

In ***Monte Carlo cross-validation*** the original training set is split randomly into training and test set. The process can be repeated many times. The advantage of this method (compared to k-fold cross validation) is that the proportion between training and test sets does not depend on the number of repetitions in the internal validation procedure.

In ***bootstrapping*** the new training sets is populated from the original training set by random sampling with replacement until the size of the new training set reaches the size of the original training set. The data not selected for the new training set becomes the new test set. On average, about 63% of original training set data goes into the new training set (some of the data appear more than once) and 37% remains in the new test set. The process can be repeated many times. One of the advantages of this method is the fact that the new training sets and the original training set are equally sized.

Monte Carlo cross-validation and *bootstrapping* suppose creating a large number of new training/test sets (1,000 – 10,000).

Alerts in OASIS models for genotoxicity

OASIS models for genotoxicity produce predictions based on alerts. Alert is a structural fragment which, if found in predicted chemical or in some of its metabolites, serves as an indicator for presence of effect (the predicted chemical is "positive" – i.e. it causes the effect predicted by the model). The different alerts act independently, which means that one chemical could be predicted positive by more than one alert. The purpose of any alert is to select only positive chemicals among the population. It is not expected one alert to select all positive chemicals, but it is desired all chemicals selected by an alert to be positive.

The usefulness (performance) of an alert is measured by the proportion of positive chemicals within all chemicals selected by the alert. The alert performance has a meaning of a probability and is fitted by a ***beta distribution***. An alert could be considered useful if its performance is statistically higher than the proportion of positive chemicals within the whole training set.

Assessing the performance of alerts

Each alert can be tested against the data from training set and its performance can be compared with the performance of a naïve alert.

A ***naïve alert*** is an alert which has no ability to select positive chemicals, i.e. its performance is equal to the proportion of positive chemicals in the training set.

The ***alert performance*** is assessed by the proportion of positive chemicals within all chemicals selected by the alert. As the real performance is not known, we can describe it via its probability distribution which is a special statistical function – ***beta distribution***. The *performance estimation* is equal to the expectation of the beta distribution inferred from the pair [*correct; incorrect*] applications of the alert.

NOTE:

In all following explanations ***alert performance*** *will stand for* ***alert performance estimation*** *calculated from the sample [correct; incorrect] applications.*

Although the alert performance shows what is the accuracy of the alert in selecting positive chemicals, this estimation lacks some important information – what is the proportion of positive chemicals in the population over which the alerts works. Indeed, it is much easier to select 5 out of 5 from a population with 70% positive chemicals than to select 5 out of 5 from a population with 30% positive chemicals. A measure that includes the latter is called ***p-value***. The *p-value* is the probability a performance equal or better than the performance of the alert to be achieved by chance if drawing chemicals randomly from training set. In this manner the *p-value* can serve as a measure for the ***reliability*** of alert performance estimation. The lower *p-value* means higher reliability and if the *p-value* cannot complement the confidence level to 1 it shows that the alert performance is statistically confirmed at the used confidence level. For example, if the confidence level is 95%, the *p-value* must be smaller than 5% in order alert performance estimation to be confirmed[1)].

Depending on their performance alerts can be classified as follows:

- ***confirmed*** – when their performance is significantly higher[1)] than the performance of a naïve alert,
- ***disproved*** – when their performance is significantly lower[1)] that the performance of a naïve alert,
- ***undecided*** - when their performance cannot be distinguished statistically[1)] from the performance of a naïve alert,
- ***theoretical*** - when they have no application over the training set.

[1)] The default confidence level used in the analysis is 95%.

NOTE:

If an alert is classified as undecided, it doesn't mean that it would have weak performance. Such classification may also appear if the number of chemicals selected by the alert is small (i.e. due to data insufficiency).

In both cases – small count of selected chemicals or performance closer to the performance of a naïve alert, – the p-value is higher than 5%.

Examples of alert performances and their classification:

Example #1 – 12 correct, 1 incorrect application(s) (*p-value* = 0.0002)

Fig. 1 - The blue line represents the performance of a naïve alert; the green alert [12; 1] has a better performance and its confidence does not include the performance of a naïve alert

Example #2 – 1 correct, 0 incorrect application(s) (*p-value* = 0.420)

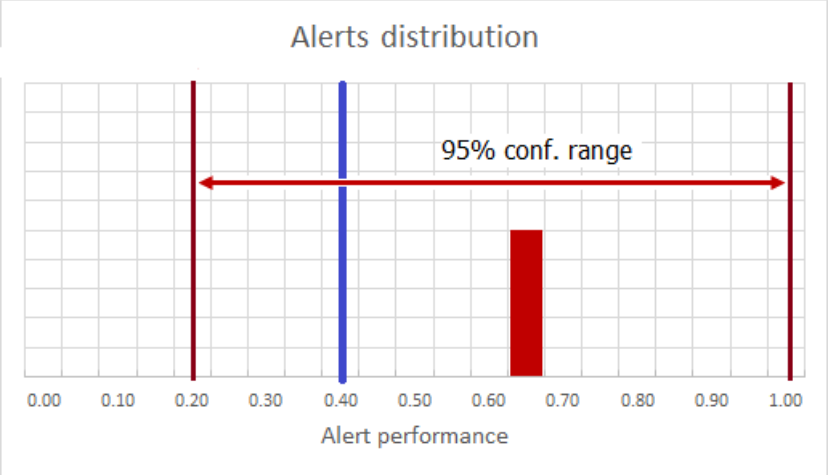


Fig. 2 - The blue line represents the performance of a naïve alert; the brown alert [1; 0] has a better performance, but it is NOT confirmed due to data insufficiency (its confidence range is too wide and includes the performance of a naïve alert)

Example #3 – 19 correct, 18 incorrect applications (*p-value* = 0.162)

Fig. 3 - The blue line represents the performance of a naïve alert; the brown alert [19; 18] has a narrower confidence range, but it is NOT confirmed because its performance is close to the performance of a naïve alert

Deriving OASIS models for genotoxicity

Each OASIS model starts with an initial library of alerts. These alerts are defined by experts and embrace existing knowledge. For any available training set these alerts could be tested against the data and their performance assessed and recorded. But along with individual alerts performances, we would be also interested in the joint performance of all alerts participating in the model.

One possible solution would be our newly derived model to use only confirmed alerts. But having in mind that many of available alerts could remain undecided due to small count of applications we prefer to do the opposite – to use all alerts (no matter confirmed or undecided) which have at least one application in the training set.

A problem could arise here if we use more alerts than needed and thus select too many chemicals (and predict them as positive). To identify such situation, we could compare the individual performances of positive and negative predictions and if the performance of positive predictions is poorer that the performance of negative predictions it would serve as an indication that the alerts used in the model are more than actually needed.

Estimation of model performances:

The performance of a model is measured in the same way as the performance of an alert – by the mean value of the beta distribution corresponding to the pair [*correct; incorrect*] applications. The reliability of the performance estimation is measured by the *p-value* for having such or better performance if drawing randomly *correct+incorrect* chemicals from the training set.

IMPORTANT NOTE:

When we talk about the performance of positive and negative predictions we have in mind the probabilities for correct predictions achieved when the model is applied on its training set. As these probabilities depend on the proportion between positive and negative chemicals in the set we try to assess, the performance values may become misleading and should be recalculated.

*Measures that do not depend on the proportion positive/negative chemicals are the performances of the model over positive chemicals (**sensitivity**) and negative chemicals (**specificity**). We provide the latter as means for calculating the performances for positive and negative predictions if the initial proportion between positive and negative chemicals is known.*

Estimation of GOF optimism:

Comparing the averaged estimations of any kind of performance (predictions performance, sensitivity, specificity) over training and test sets respectively, gives the ability to calculate the ***GOF optimism*** in this estimation and consequently to subtract it from the estimation obtained from the whole training set. Thus the corrected value is expected to be true for an external set over which our model is applied.

Given the fact that we make several types of internal validation sampling, we can average the GOF optimism values from any type of sampling and thus obtain a more reliable result.

For example, when assessing the GOF optimism (GO) for sensitivity, we can use the formula:

$$GO_{avg.} = \frac{GO_{kfold\ x4} + GO_{kfold\ x10} + GO_{MC\ 63\%} + GO_{MC\ 75\%} + GO_{bootstr.}}{5}$$

and then to calculate the sensitivity of our model over external sets:

$$Sensitivity_{ext.} = Sensitivity_{train.} - GO_{avg.}$$

The technique described above relies on results from internal validation and **must not be confused with external validation**!!

The formulas for calculating the above measures are given below:

$$Performance_{est.} = \frac{T+1}{T+F+2}$$

where

T – count of correct applications/predictions of alert/model

F – count of incorrect applications/predictions of alert/model

$$p\ value = \sum_{i=0}^{F} \left(\frac{A!}{(T+i)!\,(F-i)!} * \frac{(A'+1)!}{T'!\,F'!} * \frac{(T+i+T')!\,(F-i+F')!}{(A+A'+1)!} \right)$$

where

T – count of correct applications/predictions of alert/model

F – count of incorrect applications/predictions of alert/model

A – count of all applications/predictions of alert/model, $A = T + F$

T' – count of correct applications/predictions in the whole training set

F' – count of incorrect applications/predictions in the whole training set

A' – count of all applications/predictions in the whole training set, $A' = T' + F'$

Training set performances:

$$Sensitivity_{est.}, (performance\ over\ positive\ chemicals) = \frac{TP+1}{TP+FN+2}$$

$$Specificity_{est.}, (performance\ over\ negative\ chemicals) = \frac{TN+1}{TN+FP+2}$$

$$Performance_{est.}, all\ predictions = \frac{TP+TN+1}{TP+FP+TN+FN+2}$$

$$Performance_{est.}, positive\ predictions = \frac{TP + 1}{TP + FP + 2}$$

$$Performance_{est.}, negative\ predictions = \frac{TN + 1}{TN + FN + 2}$$

where

TP (true positives) – count of correct positive predictions

FP (false positives) – count of incorrect positive predictions

TN (true negatives) – count of correct negative predictions

FN (false negatives) – count of incorrect negative predictions

Recalculated performances over external sets:

$$Performance_{ext.}, all\ predictions = \pi * SE + (1 - \pi) * SP$$

$$Performance_{ext.}, positive\ predictions = \frac{\pi * SE}{\pi * SE + (1 - \pi) * (1 - SP)}$$

$$Performance_{ext.}, negative\ predictions = \frac{(1 - \pi) * SP}{(1 - \pi) * SP + \pi * (1 - SE)}$$

where

π – proportion of positive chemicals in the external set

SE – sensitivity (performance of model over positive chemicals)

SP – specificity (performance of model over negative chemicals)

Results of the TIMES_Ames (+S9) model

Alerts library

The model contains 93 alerts defined by experts.

Available data

The training set of the model contains 3,569 chemicals[1)], including 1,497 positive and 2,072 negative. The performance of a naïve alert is 0.420 (0.403 ÷ 0.436)[2)].

[1)] 3 inorganic chemicals were excluded, because the model has no predictions for them

[2)] Confidence range is calculated at 95% confidence level

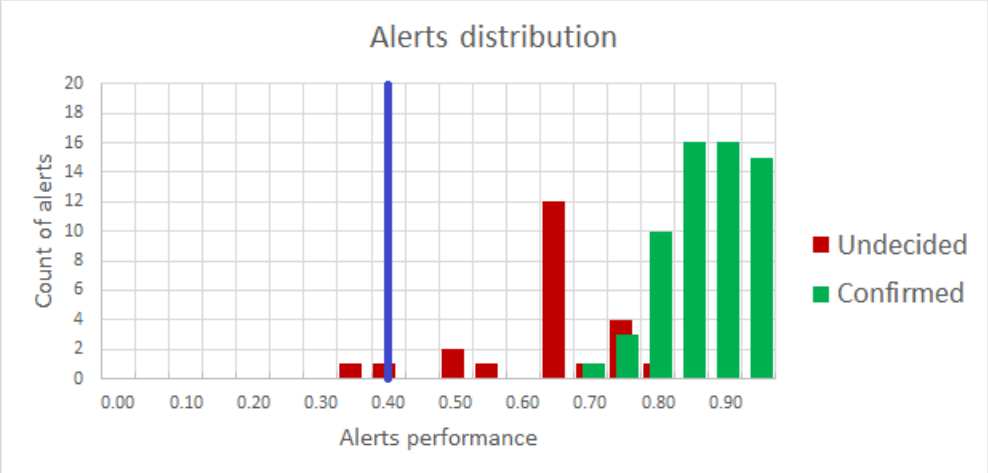
Alerts performance

The alerts performance over the whole training set is as follows:

Classification	**Count**
Confirmed	61
Disproved	0
Undecided	23
Theoretical	9[1)]

[1)] 9 alerts have no application in the training set

The distribution of alerts performance over the whole training set is as follows:



The blue line represents the performance of a naïve alert

Undecided alerts

The distribution of undecided alerts is as follows:

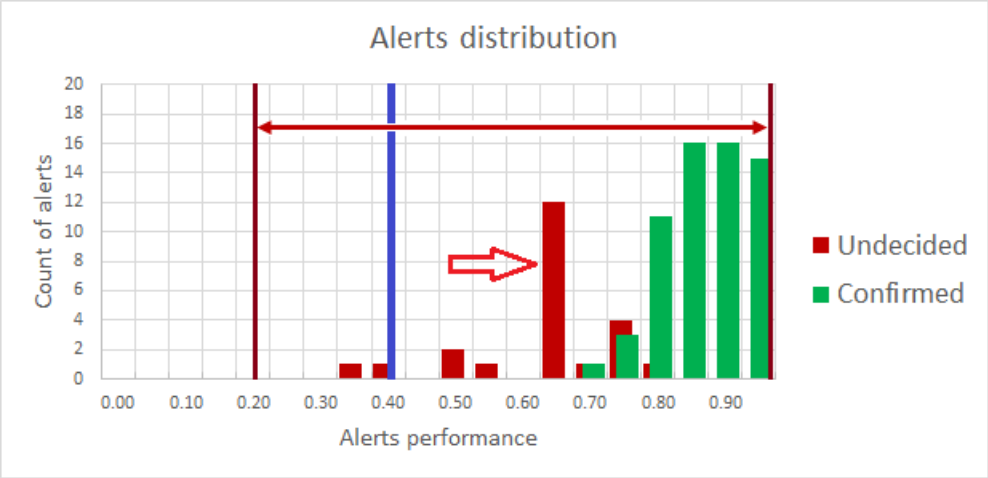
Count of applications	Count of alerts
> 10	2[1)]
5 - 9	3
3 - 4	4
2	4
1	10[2)]

[1]) "*Dicarbonyl Compounds*" and "*Quinoline Derivatives*"; although they are better than a naïve alert their performance is weak

[2]) Most of undecided alerts have only one application

The "*Dicarbonyl Compounds*" and "*Quinoline Derivatives*" alerts:

The 10 undecided alerts with one application[1)]:



[1]) The pointed bar contains also 2 alerts with 4 applications having the same performance

Model derived from the whole training set

	Performance $_{est.}$ [1)]	***p-value*** [1)]	**External performance** [2) 3)]
All predictions (accuracy)	0.875 (0.864 ÷ 0.886)	$< 10^{-10}$	0.872
Positive chemicals (sensitivity)	0.829 (0.810 ÷ 0.848)	$< 10^{-10}$	0.822
Negative chemicals (specificity)	0.908 (0.895 ÷ 0.920)	$< 10^{-10}$	0.907

[1]) Confidence ranges and *p-value* are calculated at 95% confidence level

[2]) Estimated performance for training set minus GOF optimism calculated from internal validation

[3]) Estimation of expected performance over external sets (different from training sets)

Results from k-fold cross-validation:

	10-fold		**4-fold**	
	Training sets	**Test sets**	**Training sets**	**Test sets**
Unique chemicals, %	90.0 (90.0 ÷ 90.0)	10.0 (10.0 ÷ 10.0)	75.0 (75.0 ÷ 75.0)	25.0 (25.0 ÷ 25.0)

Performance $_{est.}$, all predictions (accuracy)	0.875 (0.870 ÷ 0.880)	0.871 (0.830 ÷ 0.912)	0.875 (0.871 ÷ 0.879)	0.873 (0.858 ÷ 0.888)
p-value, accuracy	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$
Performance $_{est.}$, positive chemicals (sensitivity)	0.829 (0.820 ÷ 0.838)	0.820 (0.740 ÷ 0.900)	0.829 (0.821 ÷ 0.838)	0.823 (0.795 ÷ 0.850)
p-value, sensitivity	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$
Performance $_{est.}$, negative chemicals (specificity)	0.908 (0.902 ÷ 0.913)	0.905 (0.856 ÷ 0.953)	0.908 (0.905 ÷ 0.911)	0.907 (0.898 ÷ 0.915)
p-value, specificity	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$

[1]) Confidence ranges and *p-value* are calculated at 95% confidence level

Results from Monte Carlo cross-validation (1,000 repetitions):

	75% training set		**63% training set**	
	Training sets	**Test sets**	**Training sets**	**Test sets**
Unique chemicals, %	75.0 (75.0 ÷ 75.0)	25.0 (25.0 ÷ 25.0)	63.0 (63.0 ÷ 63.0)	37.0 (37.0 ÷ 37.0)
Performance $_{est.}$, all predictions (accuracy)	0.875 (0.868 ÷ 0.881)	0.873 (0.853 ÷ 0.892)	0.875 (0.867 ÷ 0.883)	0.872 (0.858 ÷ 0.887)
p-value, accuracy	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$
Performance $_{est.}$, positive chemicals (sensitivity)	0.829 (0.817 ÷ 0.840)	0.823 (0.788 ÷ 0.858)	0.829 (0.815 ÷ 0.843)	0.823 (0.797 ÷ 0.848)
p-value, sensitivity	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$
Performance $_{est.}$, negative chemicals (specificity)	0.908 (0.900 ÷ 0.915)	0.907 (0.884 ÷ 0.929)	0.908 (0.898 ÷ 0.917)	0.908 (0.891 ÷ 0.924)
p-value, specificity	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-10}$

[1]) Confidence ranges and *p-value* are calculated at 95% confidence level

Results from bootstrapping (1,000 repetitions):

	Training sets	**Test sets**
Unique chemicals, %	63.2 (62.2 ÷ 64.3)	36.8 (35.7 ÷ 37.8)
Performance $_{est.}$, all predictions (accuracy)	0.875 (0.864 ÷ 0.886)	0.872 (0.858 ÷ 0.887)
p-value, accuracy	$< 10^{-10}$	$< 10^{-10}$
Performance $_{est.}$, positive chemicals (sensitivity)	0.829 (0.810 ÷ 0.849)	0.822 (0.798 ÷ 0.847)
p-value, sensitivity	$< 10^{-10}$	$< 10^{-10}$
Performance $_{est.}$, negative chemicals (specificity)	0.908 (0.895 ÷ 0.920)	0.908 (0.891 ÷ 0.924)
p-value, specificity	$< 10^{-10}$	$< 10^{-10}$

[1]) Confidence ranges and *p-value* are calculated at 95% confidence level

Conclusions

The first evident observation from above results is that averaged estimations are **practically unchangeable**, no matter what kind of sampling we use for the internal validation. The variability of averaged performances is 0.003 and below even for test sets. Also their *p-values* are extremely low, which show **very high reliability** of these estimations.

The difference between performances of training and test sets - which is a measure for optimism in goodness-of-fit, - is around 0.003 for all predictions, 0.007 for positive chemicals (sensitivity) and 0.001 for negative chemicals (specificity). These values are very low and show that the model is **very well balanced** and provides **high quality** for both positive and negative chemicals. A similar quality is also expected for predictions of external chemicals.