

Exploratory Data Mining and Data Requirements to Support ToxCast Goals

Abstract

This will present an initial "broad sweep" extraction of relationships between in vitro and in vivo measurements using association rule mining principles, with the goal of identifying embedded associations and causality relationships that can guide chemical testing prioritization.

The poster will also make suggestions for more robust databases that will be eventually needed to support the ToxCast goals. The in vivo database needs to include detailed information from each animal study (e.g., chemical, endpoint, dose route, doses, numbers of animals, numbers of responders) in a standard format. An example of such a database is provided by the animal database used to compare carcinogenic potencies obtained from animal and human data (Allen et al. 1988).

Similarly, the in vivo database should include the raw measurements (e.g., well-specific responses) with systematic description of full details of experimental design (e.g., temporal and spatial arrangements of wells and plates). This amount of detail is needed to allow flexibility in definitions of response indicators, to quantify the statistical uncertainty in the estimates, and to account for any spatial and/or temporal effects in the calculation of the indicators.

Data Preprocessing :

There are three files of data ClaViv.csv, VitViv.csv, and ClaVitViv.csv. Each file is an NxP matrix, where N=321 and the value of P depends on the number of in-vitro responses and on the number of chemical property responses. The first row gives the names of chemical properties, the in-vitro response properties, or the in-vivo response properties. Each of the N-1 rows (the 1st row contains the names of columns) represent an experiment, and every column in this row represents the corresponding in-vivo, in-vitro, or chemical property response. In any case, the last 76 values in a file with in-vivo values are the in-vivo responses of either the chemical properties or the in-vivo properties. Every column has two values which are each a 0 or a 1.

The current data is very big; there are some properties for which the in-vivo response is not known. In these cases, the NA value is the response. The values in each column need to be unique to distinguish the same value from different columns. For example 0 in column one needs to be distinguished from 0 in column two and 0 in column three. This data needs to be preprocessed so that it can be used for association rule mining. To eliminate the NA values, we delete all the rows which have NA values, as they do not provide any information for prediction. Then, we change the values in a matrix (say A) to make them unique in each column by using the following pseudo code:

```
For each column i, i<=P
  For each row j, j<=N
    if A(i,j)=0
      newA(i,j)=2(i)-2
    else A(i,j)=1
      newA(i,j)=2(i)-1
    end
  end
end
```

	C1	C2	C3	...	CP
R1	1	1	1	...	1
R2	1	0	1	...	1
...	0	0	1	...	0
RN	0	0	0	...	0

Figure 1. Data Preprocessing.

Figure 1 explains this data transformation process. This dataset is then further divided into an in-vitro dataset and a rest dataset. The last 76 columns are taken from the original file, and a separate data matrix called In-VitroData is made from this dataset. The remaining columns form the second data matrix called the RestData Matrix. The size of in-vitro data is Nx76 (because the last 76 values are in-vivo values), and the size of RestData Matrix is Nx(P-76). To reduce the dimensionality 10 columns from RestData Matrix are combined with 10 columns from In-VitroData Matrix to make small datasets, each of size Nx20, where the first 10 column are the RestData values and the next 10 columns have the corresponding in-vivo values. We make all possible combinations of the RestData and In-VitroData.

Association Rule Generation:

Association rules are useful in finding causal relationships between items occurring together. They were first introduced for market-basket analysis to predict rules of the type {milk, bread} → {butter}, which, in our example, means that if a customer buys milk and bread, then he/she will most likely buy butter, too. Association rules are commonly associated with two measures of interestingness called Support and Confidence. As an example, let D be a database with N transactions. Let each transaction be a set of items from a global itemset I such that. Therefore, an association rule will be an implication, such as with support x and confidence y where support and confidence are calculated as:

$$\text{Support}(i_2, i_3 \Rightarrow i_5) = P((i_2 \cup i_3) \cup i_5) = \frac{\text{numtrans}(i_2, i_3, i_5)}{N}$$

$$\text{Confidence}(i_2, i_3 \Rightarrow i_5) = P(i_5 | i_2, i_3) = \frac{\text{numtrans}(i_2, i_3, i_5)}{\text{numtrans}(i_2, i_3)}$$

where numtrans(x) is the number of transactions in database D having x items.

It is our belief that there are some causal relationships between the in-vitro response and its corresponding in-vivo response and between the chemical property values and the corresponding in-vivo response. To extract these causal relationships we use association rule mining.

After the preprocessing step we have all the possible combinations of RestData and In-VitroData. Then, we extract association rules from these individual files with support=90% and confidence=90%. We get association rules of the form 296, 300, 341 → 1100, 1116. The rule can be transformed to imply that if the value of 149th column (represents a property) is 0, the value of 151st column is 0 and the value of 170th column is 1, the value of the 550th and 558th columns are 0. The final interpretation of rule is: if BSK_LPS_TissueFactor=0, BSK_LPS_VCAM1=0, BSK_SM3C_SAA=0 then the value of DEV_Rabbit_NeuroSensory_Brain=0 and DEV_Rabbit_Skeletal_cranial=0.

This rule is transformed for analysis using the following formula:

$$i = \frac{k}{2} + 1 \{ \forall k \text{ even} \}$$

$$i = \text{floor}(\frac{k}{2}) + 1 \{ \forall k \text{ odd} \}$$

where i is the column number and k is the current value e.g. 296, 300 or 341 where= Total number of possible values present in the original matrix.

Classifier Training and Testing:

Once the rules have been extracted for each file the next step is to combine the rules from each file into a global rule set for each type e.g. ClaViv or VitViv. The following procedure is performed:

- Take rules from each file and only keep those rules which have an in-vitro or chemical value response on the left hand side (LHS) of the rule and in-vivo response values on the right hand side (RHS) of the rule. This gives us rules predicting the in-vivo response corresponding to a particular in-vitro or chemical value.
- Assuming there are M rules in total and, we combine and into a new rule iff:

$$\left| K_{lm_j} - K_{lm_p} \right| \neq 1, \forall j, p \leq (1 \times m)$$

This ensures that the rules of the form 296, 297, 341 → 1100 are not formed as 296 and 297 are values 0 and 1 from column 149 (in other words rules of the form BSK_LPS_TissueFactor=0, BSK_LPS_TissueFactor=1, BSK_SM3C_SAA=1 → DEV_Rabbit_NeuroSensory_Brain=0 should not be formed because at a given time value of BSK_LPS_TissueFactor can either be 0 or 1). Hence, they should form different rules. After combining two rules 296, 341 → 1100 and 296, 380, 450 → 1200, we get a new rule 296, 341, 380, 450 → 1100, 1200 which translates to BSK_LPS_TissueFactor=0, BSK_SM3C_SAA=1, CLM_MitoMass_72hr=0, CLZD_CYP1A2_48=0 → EV_Rabbit_NeuroSensory_Brain=0 and MGR_Rat_ViabilityPND4=0.

Dexter O. Cahoy, Kenny Crump, and Sumeet Dua

Louisiana Tech University, Ruston, Louisiana

Association Rule Generation: cont'd

3) The combination creates rules with an LHS that has in-vitro or chemical response values and the RHS has corresponding in-vivo values. These rules can be used for classification.

4) For classification, take each instance from the classification matrix separately and divide this vector into two subvectors, one with only 76 columns having the in-vivo response values and the other with rest of the data values. Match the LHS of each rule in the global rule set with the values from the RestData subvector and check to see if their corresponding in-vivo values are the same as predicted by the rule. The current combination of rules does not predict all 76 values of in-vivo response. Therefore, we cannot make a full comparison. However, we can make a comparison using some of the predicted values that we have from the combined rules. For example, if all the combinations of our rules can predict only 20 of the 76 values for the in-vivo response, then we make a comparison with only these 20 values. For each test instance we match only these 20 predicted values (from rules) and check to see if all these values match the actual values for the same properties in the test instance. If they match, then we call this a 'Complete' (100%) match. If they do not match, then we check to see how many of these 20 values match the in-vivo values of the test instance, and we call this a 'Partial' match. This gives us the degree of match for the current instance based on the predicted values of the rules. So with our methodology not only do we provide the information of whether there is a match or not, but we also provide the degree of match. The following example explains it properly: Take for example these two test instances:

I1= {296, 340, 430, 500, 650, 1100, 1130, 1180, 1220}
I2= {296, 380, 430, 600, 650, 1100, 1120, 1190, 1220}

Here we assume the in-vivo response starts from 1100. The instances can be broken down into subvectors:

= {296, 340, 430, 500, 650} and = {1100, 1130, 1180, 1200}
= {296, 380, 430, 600, 650} and = {1100, 1120, 1190, 1220}

A combination of rules 296, 430, 650 → 1100, 1130, 1200 will have a 100% match with Instance 1, but will have a 1/3rd match with Instance 2 (because only value 1100 from the rule RHS matches the current instance's in-vivo response).

Results:

We used 70% data for training and 30% data for testing, giving 48 instances as the test subjects. We provide accuracies at two levels: one is the overall accuracy depicting how many of the test instances have a complete match, and the other depicts the percentage of partial match for each test instance. For the chemical property vs. the in-vivo response case, the overall accuracy was 72.92% (35 of the 48 instances had complete match). For in-vitro vs. in-vivo response case, the overall accuracy was 75% (36 of the 48 instances matched completely). Figure 2 shows the partial match results for both the cases. As can be seen, the partial match in all cases is above 50%.

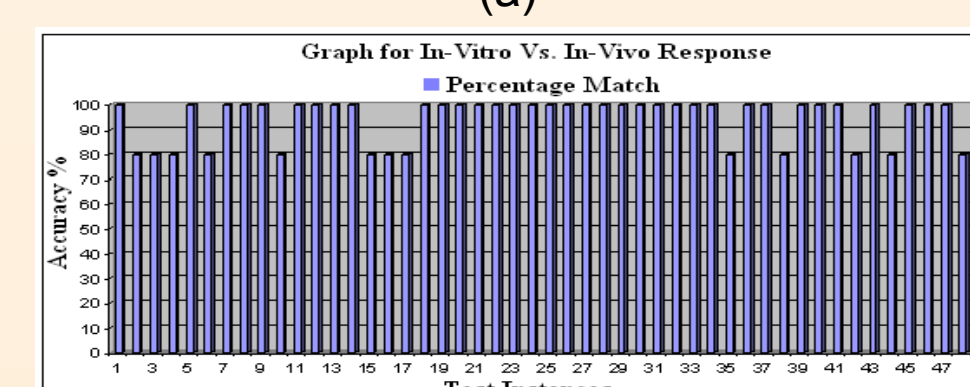
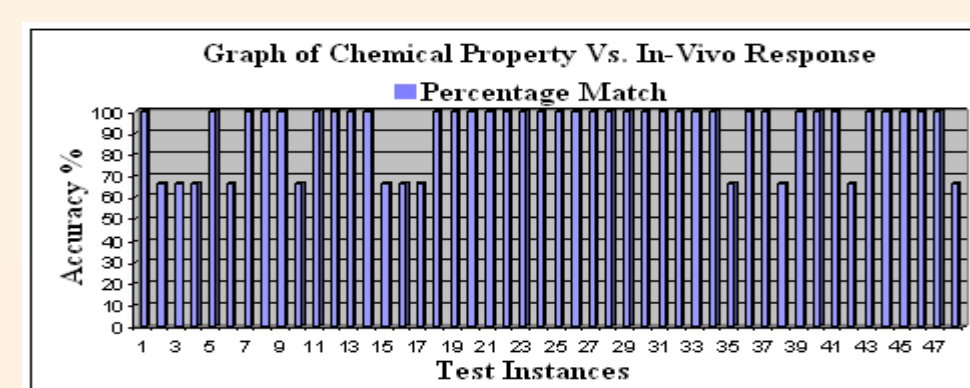


Figure 2. (a) Percentage Match of Chemical Prop. vs. In-Vivo (b) Percentage Match of In-Vitro vs. In-Vivo.

Results: cont'd

The top five rules according to confidence and their corresponding interpretation are:

8, 18, 20, 32 → 1058:

ACEA_LOC5 =0, ATG_AP_2_CIS =0, ATG_AR_TRANS =0,
ATG_DR4_LXR_CIS =0 → CHR_Mouse_LungTumors=0.

56, 58, 60, 62 → 1058:

ATG_FXR_TRANS=0, ATG_GATA_CIS=0, ATG_GLI_CIS=0,
ATG_GRE_CIS=0 → CHR_Mouse_LungTumors=0.

4, 356, 386, 404 → 1078:

ACEA_LOC3=0, CLM_CellLoss_1hr=0, CLM_MitoticArrest_1hr=0,
CLM_p53Act_1hr=0 → CHR_Rat_TesticularAtrophy=0.

68, 124, 194 → 1092:

ATG_HNF4a_TRANS=0, ATG_PPARG_TRANS=0, 3C_TissueFactor=0
→ DEV_Rabbit_Cardiovascular_MajorVessels=0.

6, 24, 64 → 1104:

ACEA_LOC4=0, ATG_CAR_TRANS=0, ATG_GR_TRANS=0 →
DEV_Rabbit_Orofacial_JawHyoid=0.

Further we also check to see which values are more prevalent in the rule RHS. This gives us an understanding of which in-vivo values play major role for classification purposes. Figure 3 [(a) and (b)] shows the count for values in rule RHS for both cases.

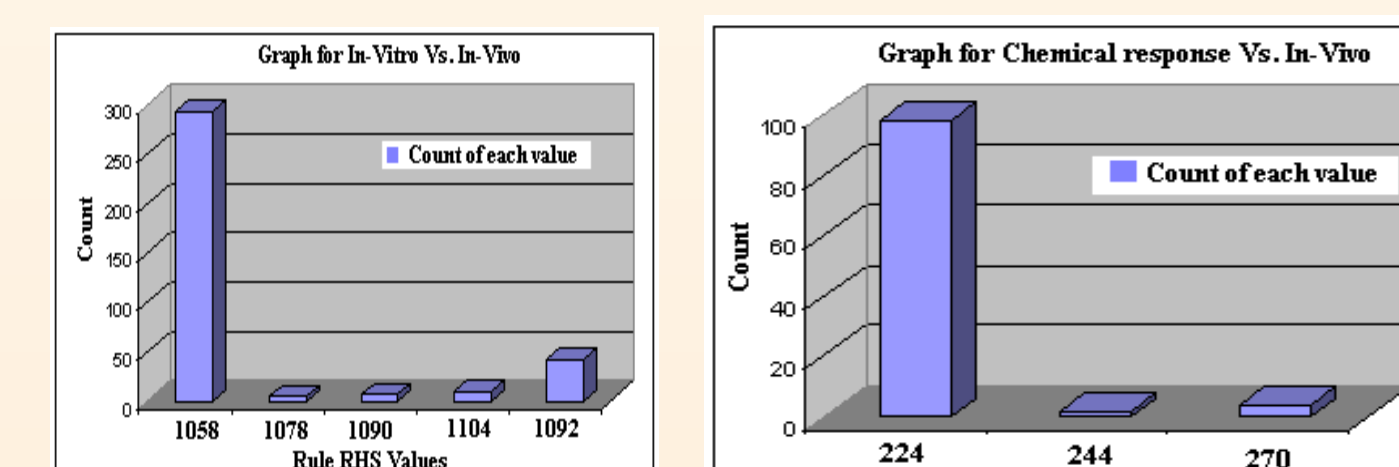


Figure 3. (a) Count for RHS Values of In-Vitro vs. In-Vivo (b) Count for RHS Values of Chemical Property vs. In-Vivo Response.

As can be seen from the graphs, this work is of value, because, in both cases, the values appear on the rule RHS most of the time. Further analysis of data is needed to ascertain this behavior.

Limitations and Future Directions:

- The current data set is binary with only 0's and 1's, which is the reason we do not get varied association rules. Proper discretization of data should be used in-order to get more interesting rules.
- Due to the current data nature, the support was kept high (90%) for rule generation which still resulted in close to 700, 000 rules for some cases, and not all these rules were useful. Proper discretization will allow the user to generate fewer and more interesting rules with lower support values.
- The current Apriori algorithm cannot generate constraint-based rules, and post-pruning of rules is performed to get rules where RHS is an in-vivo value. A better constraint-based algorithm will be used in the future to generate the rules and save time.
- The current methodology only helps in classification, but in the future we will also perform prediction and the values which are not filled by associations rules will be filled using the pairwise hamming distance between the test instance and the training instances. This will help us to get a complete response to an unknown in-vivo instance.

Improving the ToxCast Database:

The current ToxCast database is quite extensive, both in terms of the number of in vitro assays and number of chemicals included. However, to adequately carry out comparisons between in vitro and in vivo responses needed to support the ToxCast goals, even more robust in vivo and in vitro databases will be required. The in vitro data presently recorded in the ToxCast animal data base consist of Lowest Effect Levels (LELs) for each chemical and each in vivo endpoint. It is important to realize that these LELs are not purely characteristics of the chemicals and endpoint; they are also reflect numerous facets of the underlying experiments, including:

numbers of studies –chemicals evaluated in more studies tend to have lower LELs

size of study – studies that examine more animals tend to produce lower LELs

dose spacing – the LELs are determined in part by the doses used in the studies

statistical test – the LELs are determined in part by the statistical tests employed and the level of significance used.

Comparisons with in vitro responses will ultimately need to be based on a more robust chemical-specific parameter, e.g., a measure of the potency of the chemical in the test system. Data will need to be available for estimating such parameters and evaluating the uncertainty in those estimates. This will require a more detailed in vivo data base than is presently available. Such a data base needs to have critical information from each animal study (e.g., chemical, endpoint, dose route, doses, numbers of animals, numbers of responders) included in a standard format. An example of such a data base is provided by the animal data base used in a study to compare carcinogenic potencies obtained from animal and human data (Allen et al. 1988).

Similarly, a more detailed in vitro database is also needed. This data base should include the raw measurements (e.g., well-specific responses) with systematic description of full details of experimental design (e.g., temporal and spatial arrangements of wells and plates). Appropriate summarizations can easily be provided for researchers who do not wish to work with this level of detail. This amount of detail is needed to determine if the summary statistics (which presently are in the form of IC50, AC50, EC50, LEL computed using particular methods involving certain conventions, e.g., defining a positive effect based on being more than three standard deviations away from responses to DMSO) are appropriate, to quantify the statistical uncertainty in the estimates, to explore other summary statistics, and to adequately account for any spatial and/or temporal effects in the calculation of the summary statistic.

An example of the importance of having access to basic in vitro data is provided by Crump (2007). In this instance the basic data had been lost along with knowledge of exactly how the data had been summarized, and it was shown that different conclusions could be reached depending upon how the summarization was conducted.

Analysis of Replicates:

Each in vitro assay in TOXCAS was applied to three of the chemicals tested in triplicate and separately to five chemicals tested in duplicate. We studied the concordance among results from these replicates. Table 1 summarizes the replicate data by type of in vitro test. For example, with the 21 ACEA in vitro tests applied to the triplicate chemicals, 3 tests were all positive in all three replicates (+++), 15 were all negative (---) and 3 had a conflict ([1+ and 2-] or [2+ and 1-]). As a measure of reliability of a specific in vitro test we used the complement of the probability of getting the observed number or fewer of conflicts if the observed positive responses from an in vitro test were assigned by chance to the totality of results (10 for duplicates and 9 for triplicates).

Table 1 reports the average reliability across all in vitro tests within a class. Most of the in vitro results were negative and, to distinguish between tests with high reliability and those that were simply insensitive, the average was taken only over all in vitro assays that recorded at least one positive response. Overall, there appears to be only limited agreement among replicates. The Novascreen, Gentrionix and Solidus types of tests showed the greatest reliability, although the latter two types included only a limited number of tests. These results are relative insensitive to different definitions of "positive."

	Results for Duplicates					Results for Triplicates				
	All Pos.	Conflict	All Neg.	average reliability	n	All Pos.	Conflict	All Neg.	average reliability	n
ACEA	4	9	22	0.32	6	3	3	15	0.54	5
Attagene	20	33	352	0.34	30	31	28	184	0.44	34
BioSeek	86	75	274	0.56	72	14	77	170	0.39	56
Cellulmen	14	17	134	0.42	25	5	16	78	0.51	17
CellDirect	6	17	217	0.10	29	12	27	105	0.33	33
Gentrionix	1	0	4	0.89	1	0	0	3	NA	0
NCGC	0	0	120	NA	0	0	0	72	NA	0
Novascreen	22	11	1162	0.61	28	29	8	680	0.90	33
Solidus	4	0	16	0.89	4	4	1	7	0.94	4

References:

- Allen et al. (1988). Correlation between carcinogenic potency of chemicals in animals and humans. *Risk Analysis*, 8(4), 531-544.
- Crump, K. (2007). Letter to the editor: Limitations in the National Cancer Institute Antitumor Drug Screening Database for evaluating hormesis. *Toxicological Sciences*, 98(2), 599-601. doi: doi:10.1093/toxsci/kfm135.