Using Mutual Information to Quantify Toxicological Relationships in Quantitative High Throughput Screening Data

NIH National Institute of Environmental Health Sciences

Sankar, Akshay[1]; Shockley, Keith[2]

[1]University of North Carolina – Chapel Hill, [2]NIEHS Biostatistics and Computational Biology Branch

Abstract

BACKGROUND: Quantitative high throughput screening (qHTS) is an *in vitro* approach used to simultaneously test many chemicals over a broad range of concentrations to better assess chemical toxicity. Large qHTS datasets have several unidentified toxicological relationships. Mutual information (MI) is an information theory statistic used to describe the dependence between random variables. MI can describe both linear and non-linear relationships between chemicals in units of bits (0 or 1), where the value of MI represents the number of possible subgroups (2^{bits}) of paired responses between two chemicals. An MI of 0 bits describes a statistically independent relationship between chemicals.
OBJECTIVE: To develop an MI-based approach to describe toxicological relationships between biologically active chemicals in the BG1 ER agonist assay (phase II of Tox21).
METHODS: A total of 8,306 chemicals surveyed in the estrogen receptor agonist assay were filtered to 1,093 active chemicals. Interpolating splines were used to account for missing data and different concentration spacing. MI estimates were compared to Pearson correlation. Significance was based on permutation testing.
RESULTS: Observed MI values ranged from 0 bits to 3.9-bits across all 596,778 pairwise comparisons. In general, MI estimates captured linear and non-linear relationships between chemicals. A total of 80,333 and 19,458 significant chemical pairs (α = 5%) were found in the BG1 ER agonist assay using Pearson and MI, respectively.
CONCLUSIONS: MI can be used to quantify non-linear relationships between chemicals and find associations between chemicals not identified by Pearson correlation. Splines can be used to correct for missing data and offset concentrations ranges between chemical profiles. A scaled MI value, ranging from 0 to 1, may be useful to compare MI values of different chemical pairs.

Introduction

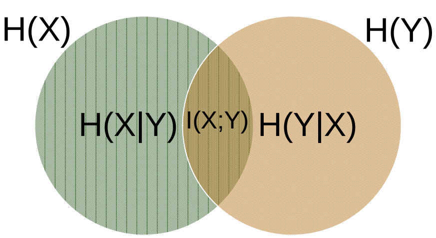
Quantitative high throughput screening (qHTS) produces concentration-response data for thousands of chemicals simultaneously and has applications in drug discovery and toxicity testing (Collins et al., 2008).

In Tox21, about 10,000 chemicals are tested at 15 concentration levels to more accurately assess toxicological response. qHTS datasets will be generated for 200-300 different assays (Shockley, 2015).

Describing toxicological relationships in qHTS data may help toxicologists prioritize chemicals for further study or, ultimately, predict toxicity.

Entropy (*H*) represents how much information (or uncertainty) is conveyed by the probability distribution of the chemical profile (e.g., chemical X or chemical Y). Mutual information (*I*) is a measure of the amount of information shared between chemical profiles (Shannon, 1948).

Mutual information is a nonparametric measure of dependence that can describe linear and non-linear toxicological relationships.



Objective

The underlying concentration-response profiles in qHTS data vary widely across different chemicals. Thus, the form of toxicological relationships between chemicals in qHTS data are not known: they may be linear, exponential, quadratic etc.

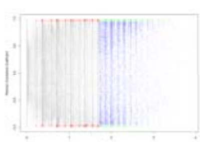
We sought to develop a nonparametric approach to:

- identify relationships between chemicals (linear or nonlinear)
- determine the significance of these associations
- compare our new measure with Pearson correlation

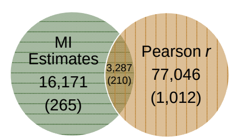
Methods

- A total of 8,306 chemicals in the BG1 ER agonist assay were filtered to 1,093 active chemicals with homogeneous profiles.
- Interpolating splines were used to estimate response values along each individual profile and account for differences in concentration spacing.
- Response values were discretized into three standard deviation intervals of the negative controls. MI estimates were computed using discretized values according to Equation (1).
- The significance of the observed MI and Pearson correlation values were obtained via permutation testing after shuffling the BG1 ER agonist data set 30 times.
- Using MI values, the largest cliques, maximal and complete subgraphs, of significant pairwise comparisons (α = 5%) were analyzed with the Tox21 Enricher tool [http://hurlab.med.und.edu/tox21enricher/] to identify over-represented KEGG pathways.

Figure 1: MI vs. Pearson Correlation

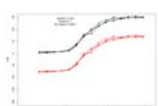


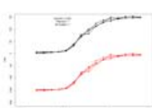
Scatterplot of all 596,778 pairwise comparisons between 1,093 chemicals with homogeneous profiles contained in the BG1 ER agonist assay qHTS dataset. A significance level of 5% was used to determine colored points on the plot. Pairwise comparisons that were not significant are show in gray, only significant Pearson coefficient values in red, only significant MI values in blue, and significant MI and Pearson coefficient values in green.



Venn diagram of significant (α = 5%) pairwise comparisons between chemicals with homogeneous profiles in the BG1 ER agonist assay qHTS dataset. The number of unique chemicals in these pairwise comparisons are shown in parentheses.

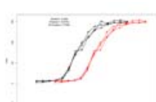
Figure 3: Shifted Profiles

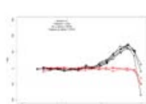




Unlike Pearson correlation, MI can be more sensitive to vertical translations of concentration - response profiles. A scaled MI value may help identify vertical translations in the data.

MI is more sensitive than Pearson correlation for horizontal translations of concentration-response curves. MI also responds to the degree of horizontal shifts of the chemical profiles.





An MI of 0-bits indicates a statistically independent relationship between the concentration-response profiles of the two chemicals. Typically, a statistically independent relationship occurs when the responses of one chemical is not informative (i.e. flat lines).

Relevant Equations

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) \quad (1)$$

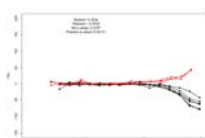
$$H(X|Y) = -\sum_{i=1}^{m}\sum_{j=1}^{n} p(x_i, y_j)\log_2 p(x_i|y_j) \quad (2)$$

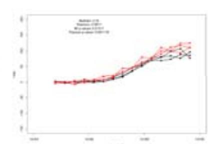
$$I(X;Y) = H(X) - H(X|Y) = \sum_{x \in X}\sum_{y \in Y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)} \quad (3)$$

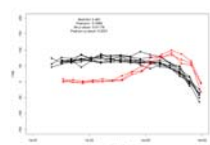
$$I(X;Y) = I(Y;X) \geq 0 \quad (4)$$

- The mutual information, I(X;Y), is the reduction in entropy (or uncertainty) of one variable (e.g., one chemical) given observations of the other variable (or a different chemical).
- Mutual information increases monotonically, is symmetric and is 0 when the random variables are statistically independent.
- An MI of 0 implies the random variables are statistically independent [P(x,y) = P(x) P(y)].

Figure 2: Experimental qHTS Examples

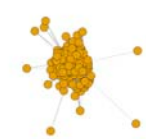






Example concentration-response profiles of two chemicals (red and black) with insignificant MI and significant Pearson correlation (top left), significant MI and insignificant Pearson correlation (bottom left), and significant MI and Pearson correlation (top right) at the 5% significance level. Non-linear associations are captured by the MI-based approach.

Figure 4: Tox21 Enrichment Analysis



Undirected graph showing nodes (chemicals) and edges (associations) of significant (α = 5%) pairwise comparisons of chemicals in the BG1 ER agonist data using an MI-based approach. One cluster of associated chemicals was found. Several maximal cliques of various sizes were found. Cliques are collections of vertices (subgraphs) that are complete, i.e. each possible edge is present. A maximal clique is a complete subgraph that does not extend to a larger complete subgraph.

	Term	Count	PValue	Benjamini (False Discovery Rate)
Clique 1	KEGG:00140|Steroid hormone biosynthesis	21	1.94E-08	4.01E-06
	KEGG:04114|Oocyte meiosis	15	2.48E-05	0.002563615
	KEGG:01100|Metabolic pathways	22	3.02E-05	0.00208618
	KEGG:00500|Starch and sucrose metabolism	10	6.08E-05	0.003148238
	KEGG:04914|Progesterone-mediated oocyte maturation	17	7.52E-05	0.003112342
	KEGG:05217|Basal cell carcinoma	7	0.000156067	0.005384296
	KEGG:04961|Endocrine and other factor-regulated calcium reabsorption	7	0.000274228	0.008109307
	KEGG:00860|Porphyrin and chlorophyll metabolism	10	0.000461527	0.011942016
	KEGG:02010|ABC transporters	9	0.000804646	0.018506857
	KEGG:00514|Other types of O-glycan biosynthesis	8	0.000812094	0.016810339

Enrichment analysis, showing significant KEGG pathways, of the largest cliques (4 cliques of 115 chemicals) in the graph shown above. All four cliques contained the same set of significant KEGG pathways: most notably, steroid hormone biosynthesis, progesterone-mediated oocyte maturation and basal cell carcinoma. Clique 1 is shown as an example. Enrichment conducted using Tox21 Enricher (http://hurlab.med.und.edu/tox21enricher/).

Conclusions

The MI-based approach:

- Captures about 16,000 associations between chemicals
- Accounts for missing data and differences in concentration spacing
- Is scale invariant and translation variant
- Nonparametric and symmetric

…with some caveats:

- MI fails to find significant associations between chemicals with uninformative (e.g. flat-line) concentration-response profiles
- Data discretization (3 standard deviations of negative controls) may not be optimum

Assessing the significance of MI estimates is difficult:

- Different pairwise comparisons might have different possible maximal MI values
- No probability distributions are available to reliably characterize the significance of small *and* large MI values
- P-value correction for the multiple testing problem was not addressed

Future Directions

Q1: Can we use Markov models to relax the independent and identically distribution (*iid*) assumption and more accurately measure MI?

Q2: Can we optimize discretization procedures to maximize the mutual information between two chemical profiles?

Q3: Can we develop methods to reliably compare and assess the significance of MI observations?

Q4: How can MI be extended to describe toxicological relationships between three or more chemicals?

References

1. Collins FS, Gray GG and Bucher JR. (2008). Transforming environmental health protection. Science. 319: 906-907.
2. Shockley KR. (2015). Quantitative high-throughput screening data analysis: challenges and recent advances. Drug Discovery Today 20: 296-300.
3. Shannon CE. (1948). A mathematical theory of communication. Bell System Technical Journal. 27: 379–423, 623–656.

Acknowledgments

Summer Internship Program (NIEHS)

Branch of Computational Biology and Biostatistics (NIEHS)

Tox21 (NTP) for BG1 ER agonist data: https://tripod.nih.gov/tox21