Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Research Article
Volume 5 Issue 2 - February 2018
DOI: 10.19080/BBOAJ.2018.04.555660

Biostat Biometrics Open Acc J
Copyright © All rights are reserved by JB Kadane

Bayesian Genotoxicity

Keon Woo Kim[1] and JB Kadane[2]*

[1]*Department of Statistics, Pennsylvania State University, USA*

[2]*Department of Statistics, Carnegie Mellon University, USA*

Submission: January 25, 2018; **Published:** February 27, 2018

*Corresponding author: Joseph B Kadane, Department of Statistics and Data Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania, 15213, USA Email: kadane@stat.cmu.edu

Abstract

This paper reports an initial exploration relating the genotoxicity of chemicals tested in vivo to the genotoxicity of chemical structures (sets of chemicals with common components). The analysis uses genotoxic results on 568 chemicals to make inference about the genotoxicity of 2893 chemical structures. As a result, there is substantial residual uncertainty. Two simple models are proposed to relate the genotoxicity of a chemical to the genotoxicity of its component chemical structures. Using these models, we find 19 chemical structures that are among the (predicted) most toxic 15 under at least one of the two models. Of these, 11 are already known to be genotoxic; we suggest further in vivo study of the remaining 8. The method used in the paper is Bayesian Markov Chain Monte Carlo. Convergence was rapid for both models. When applied predictively to chemicals, the results are disappointing: nearly every chemical is predicted to be highly toxic. To remedy this, we propose that further progress along this line would depend on the use of biochemically more sophisticated models that focus only on those chemical structures that are bioactive.

Keywords: Bayesian Analysis; Chemical Structures; Markov Chain Monte Carlo; Toxicity

Introduction

Genotoxicity

During cell division in mammals and other animals, DNA packaged into chromosomes divide as two daughter cells are created. If a chemical disturbs this process, chromosomal damage can result. Such damage can lead to cancer, birth defects, fetal deaths and infertility. There are standard methods to test a chemical to see if it induces chromosomal damage [1]. However, such tests are expensive and take time, which has led to interest in using in silico models to predict genotoxicity. The purpose of this paper is to examine the extent to which genotoxicity results can be associated to chemical substructures (i.e. structural features), sets of chemicals that share parts of chemicals with DNA and that might react similarly. To this end, 568 chemicals have been tested for their effects on DNA, and are characterized by 2893 structural features.

Data sources

There were two sets of structural feature sets used to annotate the chemicals, the default set of features and the genetox alerts. The reason there is redundancy in these is the same structural features can occur in the different feature sets. This was done to be comprehensive in annotating the chemical substructures. The *in vivo* mouse micronucleus results for the chemicals in the Tox21 10K library were derived from the in silico first tool (https//www.leadscope.com/isfcui/app#). An sdf file provided by the US Environmental Protection Agency through their Tox21 web portal (http://www.epa.gov/comptox/dsstox/sdf_tox21s.html) was loaded into the insilico first application and the in vivo mouse micronucleus activity annotations, derived from the Lead scope SAR Genotoxicity Database (http://www.leadscope.com/toxicity_ databases/), were mapped to the chemicals. In total 575 of the chemicals in Tox21 library were annotated with in vivo mouse micronucleus (139 positives and 436 negatives).

Chemical structures were retrieved from the Environmental Protection Agency's (EPA) annotation of the Tox21 10K library (http:/www.epa.gov/comptox/dsstox/sdf_tox21s.html). The SDF file provided by the EPA was loaded into the Lead scope software (Version 3.1; Columbus, OH). For the 575 chemicals with micronucleus data a total of 2893 structural features representing medicinal chemistry building blocks were generated. Of the 575 chemicals with micronucleus results 568 contained at least one structural feature. The remaining 7 without a single structural feature were removed from the data set and not considered further. Of the 2893 structural features 2797 are from the default Lead scope structural feature set (identifier F2 -27070). An additional 96 structural features (F27143¬F27287) related specifically to DNA damage were also included in order to ensure comprehensive coverage of the structural features most relevant to the biological endpoint under consideration. Some of the DNA damage specific structural features were redundant with those contained in the default set of structural features, however if the identical features occurred in both sets they were assigned different identifier numbers and were left in the data set, hence the same structural feature may be associated twice.

Methods

There are two kinds of data that enter the analysis. The primary data specify, for each chemical, whether it was found to be genotoxic. The secondary data specify, for each chemical i, the structures it contains. The analysis then permits inferences about the genotoxicity of each chemical structure, taking both sources of information into account. Making inferences about the genotoxicity of 2893 chemical structures on the basis of the known genotoxicity of 568 chemicals inevitably involves substantial residual uncertainty. To make the link between the toxicity of a chemical and the toxicity of its constituent chemical structures requires a model. There are two models that are used for the purpose of this research. Let S_i denote the set of chemicals sharing chemical structure $i(i=1,\ldots,2893)$. Suppose chemical structure i has probability p_i of not being genotoxic. Then what is the probability q_j that chemical j is benign (i.e., not toxic)?

In model I, we suppose that

$$q_j = \prod_{j \in S_i} p_i \qquad (1)$$

In words, a chemical is benign if each of the structures to which it belongs, acting independently, is benign. As a consequence, the probability that chemical j is genotoxic is

$$1 - q_j = 1 - \prod_{j \in S_i} p_i . \qquad (2)$$

In model II, we suppose that the probability that chemical j is genotoxic is

$$1 - q_j = \max_{j \in S_i} \left(1 - p_i\right), \quad (3)$$

the largest of the probabilities that each constituent chemical structure is genotoxic. Then, under model II, the probability that the chemical structure j is benign is

$$q_j = 1 - \max_{j \in S_i} \left(1 - p_i\right) = \min_{j \in S_i} p_i . \qquad (4)$$

It is likely that toxicologists have prior information about the suspected genotoxicity of at least some chemical structures. However, to elicit such information about each of the 2893 chemical structures in question would probably exhaust the patience of both toxicologists and statisticians. Consequently, as an initial computation, we take the prior distribution on $p = (p_1, \ldots, p_{2893})$ to be independent and uniform on (0, 1).

Results

Each model was run for 5000 iterations of a Metropolis-Hastings MCMC algorithm [2]. Index plots suggest immediate achievement of equilibrium, so we used all 5000 sampled probabilities. As a check, we ran a second version that differed from the first in the following way: if chemical structure j's initial probability in the first run was q_j^0, the second run started at $1-q_j^0$ for each j. Figures 1 & 2 plot the means against each other for these two runs. The results reinforce the conclusion that the MCMC's have reached equilibrium. On average, the 568 chemicals tested have about 30 chemical structures each. Consequently, applying Model I would predict that virtually every chemical is genotoxic. Model II is only slightly better in this predictive sense. Improving these models would require a much more detailed understanding of how chemical structures contribute to the genotoxicity of chemicals than we now have. Nonetheless, and despite the large standard errors in our estimates, we think there is value in identifying suspect chemical structures. That 11 of our 19 had already been found to be genotoxic is a kind of validation. The other eight suggest a possible route of further research. Table 1 records the fifteen most genotoxic chemical structures found under either of the two models. Lines 1-15 are the most toxic (least benign) according to Model I. Lines 1-9, 12, 13, 16-19 are the chemical structures most toxic according to Model II. None of the 19 listed chemical structures were included in chemicals tested as benign.

Table 1: Means and Standard Deviations of p, probability of being benign, under each model, and number of genotoxic chemicals, for each of 19 least benign chemical structures.

			Model I		Model II		Genotoxic Chemical
	Common Name	Structure ID	Mean	St. Dev.	Mean	St. Dev.	
1	Carbodimide	F6578	0.209	0.171	0.391	0.279	2
2	Sulfamic acid ester	F6612	0.227	0.187	0.969	0.298	1
3	Arsenic	F5253	0.297	0.193	0.26	0.206	2
9	Imidazole 2-methyl	F13008	0.25	0.213	0.446	0.301	2
5	Monohaloaldkenes	F27213	0.296	0.234	0.404	0.296	2
6	Halogenated methanes	F27151	0.296	0.23	0.429	0.293	1
7	Methane, 1, 1-dihydroxy	F25966	0.307	0.235	0.436	0.295	1
8	Propane, 1-carbonyloxy-,3-hydroxy-	F26985	0.392	0.257	0.963	0.3	1

9	Nitrate	F6586	0.353	0.294	0.926	0.292	1
10	Benzene, 1-alkenyl-,3-methoxy	F1930	0.376	0.273	0.475	0.298	2
11	Benzene, 1-amino(NH2)-,9-chloro-	F3891	0.392	0.277	0.965	0.299	3
12	Alkyl nitrate	F27228	0.399	0.278	0.961	0.297	1
13	Methane, 1,1-dicarbonyl-	F25859	0.9	0.268	0.957	0.298	1
14	Pyridine, 2-heteroamino-	F17826	0.41	0.282	0.979	0.3	2
15	Carboximid0H2), alkenyl,acyc-	F5715	0.936	0.281	0.473	0.295	1
16	Nitrosamine, including imidacloprid	F27215	0.98	0.288	0.958	0.298	3
17	Carboximid0H2), alkenyl	F5714	0.972	0.289	0.46	0.295	1
18	Nitrosamine, excluding imidacloprid	F6591	0.977	0.287	0.46	0.299	3
19	Sulfunate, 0-alkyl-	F7091	0.493	0.287	0.463	0.3	3

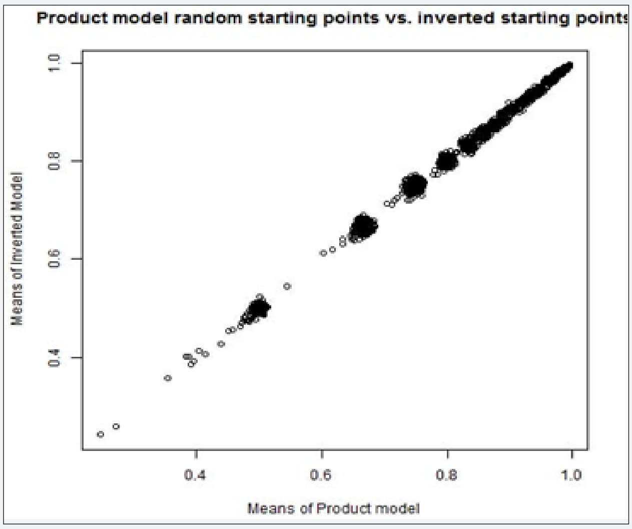

Figure 1: Relationship between probabilities of chemical structures being benign for same model but with inverted starting probabilities Model I (Product). Means are averages over 5000 replicates.

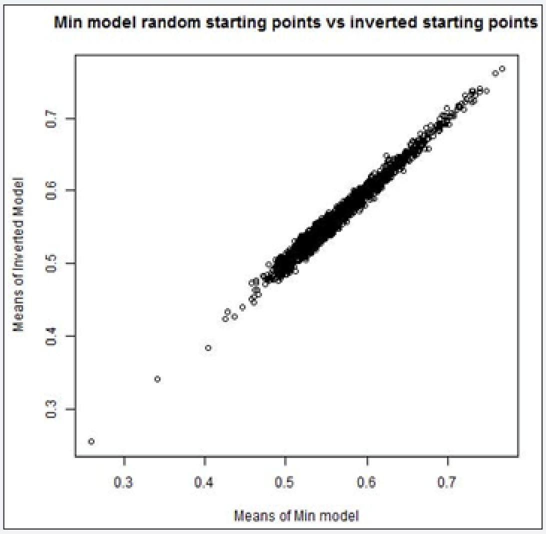

Figure 2: Relationship between probabilities of chemical structures being benign for same model but with inverted starting probabilities Model II (Min). Means are averages over 5000 replicates.

The standard deviations reported here are calculated as if the MCMC draws are independent. In view of the likely positive serial dependence among them, these calculated standard deviations should be interpreted as lower bounds. Nonetheless, they give a reasonable indication of the uncertainty inherent in the calculated means. While it was not feasible to elicit priors on all 2893 chemical structures being studied here, we report below the literature on the 19 we find most suspect, as listed in Table 1. In doing so, we omit references based on the data we are studying, to avoid double-counting. Eleven of the 19 chemical structures listed in Table 1 are mentioned in the literature as known or suspected genotoxins:

1. Carbodimide (F6578) [3]

3. Arsenic (F5253) [4]

5. Monohaloalkenes (F27213) [5]

6. Halogenated methanes (F27151) [6]

11. Benzene, 1-amino(NH2)-, 4-chloro (F3891)

12. Alkyl nitrite (F27228)

15, 17. Carboxamide(NH2), alkenyl, acyc-(F5714) and Carboxamide(NH2), alkenyl (F5715) have the strucuture ‘α, β unsaturated carbonyls,[7-9]

16, 18. Nitrosamine, including imidacloprid (F27215) and Nitrosamine, excluding imidacloprid “alkyl and aryl N-nitroso groups” (F6591) in “a well-established class of chemical carcinogens [10]

19. Sulfonate, O-alkyl-“alkyl (C< 5) or benzyl ester of sulphonic or phosphonic acid,” (F7091)

Our attitude toward the eight chemical structures on our list for which we have not found supporting literature (i.e., lines 2, 4, 7-10, 13 and 14 in Table 1) is as follows: they may be spurious, or they may be genotoxic, and we don’t know which. We would point out, however, that several of them are included in only one

How to cite this article: Kim KW, JB Kadane. Bayesian Genotoxicity. Biostat Biometrics Open Acc J. 2018; 5(2): 555660. DOI: 10.19080/BBOAJ.2018.04.555660

tested chemical, so this finding should be treated with caution. However, there is benefit in reducing the list of suspected chemical structures from the 2893 we started with, to eight. Thus, we propose that chemicals with one or more of these eight structures would be good candidates for further testing. It is also useful to get an overview of the relationship between the results for the two models. Figure 3 displays the means under Models I and II for each chemical structure, plotted against each other.

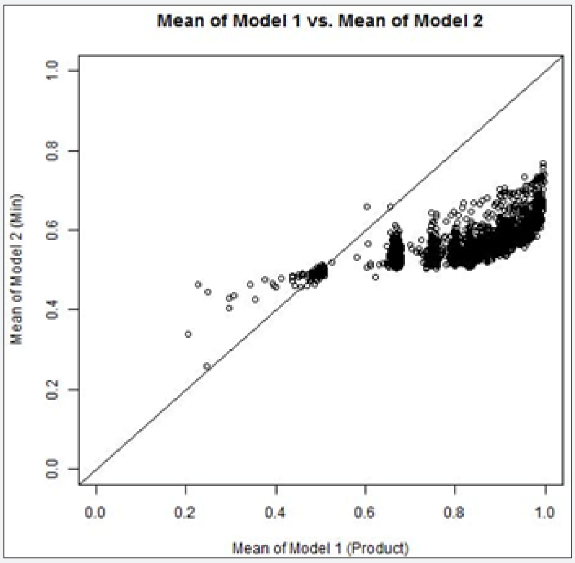

Figure 3: Relationship between the two models: Model I (Product) vs. Model II (Min), reporting the mean probability of chemical structures being benign.

Conclusion & Discussion

Strength of this analysis is that we used a standardized endpoint for evaluating genotoxicity, the micronucleus assay. The major limitation of this analysis is the poor predictivity of both models, as nearly all chemicals were predicted to be toxic. This is due to the structure of the models. A typical chemical has about 30 chemical structures. The product of 30 probabilities, as in Model I, leads to a small probability of a chemical being benign. Model II is not quite as severe, but the minimum of 30 probabilities is also likely to be small. Consequently, both models are too pessimistic in predicting the genotoxicity of, as yet untested, chemicals. The issue here is not Bayesian analysis, nor Markov chain Monte Carlo, but rather that Models I and II do not reflect enough toxicological wisdom to be useful predictively. This is not surprising given work by others that shows poor model predictively on extremely reductionist data sets used to predict apical outcomes in complex non-linear systems [11]. Because these models did not predict well, we did not pursue applying them to additional test sets to assess factors such as domain of applicability. It could be that only a few of the chemical structures in a typical chemical in the data base are bioactive. Knowing which those are would permit a much more targeted analysis that might yield better predictions. We have identified eight chemical structures, not previously thought to be genotoxic that are reasonable suspects for genotoxicity. We suggest further laboratory investigation of chemicals involving these chemical structures would be useful.

Acknowledgement

We are very grateful to Scott Auerbach, Fred Palmer and Kristina A. Thayer of the National Toxicology Program, National Institute of Environmental Health Services, Department of Health and Human Services, for their advice and support. We also thank Lead Scope for making their data available to us.

References

1. National Toxicology Program (2004) A National Toxicology Program for the 21st Century: A Roadmap for the Future. Research Triangle Park, NC: NTP, NIEHS, USA.
2. Brooks S, Gelman A, Jones G, Meng XL (2011) Handbook of Markov chain Monte Carlo. Boca Raton: Chapman & Hall, USA.
3. Moshnikova AB, Afanasyev VN, Proussakova OV, Chernyshov S, Gogvadze V et al. (2006) Cytotoxic activity of 1-ethyl-3 (3-dimethylaminopropyl) carbodimide is underlain by DNA interchain cross-linking. Cell Mol Life Sci 63(2): 229-234.
4. Bustaffa E, Stoccoro A, Bianchi F, Migliore L (2014) Genotoxic and epigenetic mechanisms in arsenic carcinogenicity. Arch Toxicol 88(5): 1043-1067.
5. Benigni R, Bossa C (2011) Mechanisms of chemical carcinogenicity and mutagenicity: A review with implications for predictive toxicology. Chem Rev 111(4): 2507-2536.
6. Morimoto K, Koizumi A (1983) Trihalomethanes induce sister chromated exchanges in human lymphocytes *in vitro* and mouse bone marrow cells *in vivo*. Environ Res 32(1): 72-79.
7. Benigni R, Giuliani A, Franke R, Gruska A (2000) Quantitative structure-activity relationships of mutagenic and carcinogenic arometic amines. Chem Rev 100(10): 3697-3714.
8. Eder E, Hoffman C, Bastian H, Deininger C, Scheckenbach S (1990) Molecular mechanisms of DNA damage Initiated by alpha, beta unsaturated carbonyl compounds as criteria for genotoxicity and mutagenicity. Environ Health Perspect 88: 99-106.
9. Kozekov ID, Nechev LV, Moseley MS, Harris CM, Rizzo CJ, et al. (2003) DNA interchain cross-links formed by acrolein and crotonaldehyde. J Am Chem Soc 125(1): 50-61.
10. Woo YT, Lai DY, Argus MF, Arcos JC (1995) Development of structure-activity relationship rules for predicing carcinogenic potential of chemicals. Toxicol Lett 79(1-3): 219-228.
11. Thomas RS, Black MB, Li L, Healy E, Chu TM, et al. (2012) A comprehensive statistical analysis of predicting *in vivo* hazard using high-throughput *in vitro* screening. Toxicol Sci 128(2): 398-417.

How to cite this article: Kim KW, JB Kadane. Bayesian Genotoxicity. Biostat Biometrics Open Acc J. 2018; 5(2): 555660.
DOI: 10.19080/BBOAJ.2018.04.555660

DOI: 10.19080/BBOAJ.2018.04.555660