Environmental Health Perspectives
Vol. 96, pp. 81-84, 1991

Mathematical Models for Exploring Different Aspects of Genotoxicity and Carcinogenicity Databases

by R. Benigni* and A. Giuliani*

One great obstacle to understanding and using the information contained in the genotoxicity and carcinogenicity databases is the very size of such databases. Their vastness makes them difficult to read; this leads to inadequate exploitation of the information, which becomes costly in terms of time, labor, and money. In its search for adequate approaches to the problem, the scientific community has, curiously, almost entirely neglected an existent series of very powerful methods of data analysis: the multivariate data analysis techniques. These methods were specifically designed for exploring large data sets. This paper presents the multivariate techniques and reports a number of applications to genotoxicity problems. These studies show how biology and mathematical modeling can be combined and how successful this combination is.

Introduction

A general problem that is common to all scientific research is how to derive the maximum available information from the observations and data relative to a given phenomenon. In biology, for example, exactly how to approach the analysis of data is a recurrent problem. It is equally pertinent for the specific problem of understanding and using the information contained in the genotoxicity and carcinogenicity databases. One great obstacle is the very size of such databases. They consist of large amounts of information; their vastness makes them difficult to read, thus obscuring the relationships they contain. This leads to inadequate exploitation of the which becomes costly in terms of time, information, labor, and money.

Until now, in the various attempts to find a method with which to overcome such problems and to better exploit the information of the databases, examining the data by eye has been combined with various more objective tools: *a*) statistical techniques have been used to analyze specific aspects; *b*) computation of frequencies and indices such as sensitivity, specificity, etc., have been used to summarize certain parts of the information; and *c*) graphical representation of histograms have been devised.

All these various approaches certainly served in the understanding of the data. However, this search for adequate tools with which to attack large databases has, curiously, almost entirely neglected an existent series of very powerful methods of data analysis: the multivariate data analysis techniques. In fact, their foundation dates back to the beginning of this century; they were specifically designed for exploring large data sets. They contain a number of essential properties: *a*) they have reached a high level of development and sophistication; *b*) they have a clear and solid theoretical base; *c*) they have a very high degree of flexibility, as is demonstrated by the fact that they have been succesfully applied in many different fields (astronomy, social sciences, psychology, biology, quantitative structure-activity relationships, etc.); and *d*) they are standardized and are commercially available in software packages for every kind of computer. For a presentation of the various multivariate methods, see Lebart et al. (*1*). Specific applications to genetic toxicity are reported in Benigni and Giuliani (*2*).

Multivariate Data Analysis Methods

The multivariate data analysis methods can be classified into two large families: methods for summarizing and visualizing the information, such as factor analysis; and automatic classification techniques, i.e., the clustering methods. The combination of the various methods in an analysis helps to "see" the data structure from various points of view.

Factor analysis operates on objects defined by a number of variables; it generates a new set of artificial variables (called factors), whose number is lower than that of the original variables, but they still represent almost all the information provided by the original set of variables. Each factor describes one of the basic effects that play a role in the phenomenon, and factor 1 represents the most important basic effect. Mathematically speaking, the factors are linear combinations of the original variables.

Cluster analysis is another multivariate technique, which identifies groups of individuals or objects that have similar characteristics. If we have a table where the objects are defined by a number of variables, or descriptors, cluster analysis places the objects that show similar profiles of variable values in the same class.

*Laboratory of Comparative Toxicology and Ecotoxicology, Istituto Superiore di Sanita', Rome, Italy.

Address reprint requests to R. Benigni, Laboratory of Comparative Toxicology and Ecotoxicology, Istituto Superiore di Sanita', Rome, Italy.

The usefulness of cluster analysis is twofold. First, it can help reduce the complexity of an analysis by breaking the population of objects into subpopulations on which to perform further analyses. Second, when used in combination with factor analysis, it helps interpret the meaning of the factors by identifying groups of objects that characterize the ends of the axes.

One of the important aspects of multivariate techniques is their ability to reorganize the information in a more easily "readable" form. The one fundamental element that makes multivariate techniques so efficient, and which should be stressed, is that the reorganization of the information is not performed according to the ideas, feelings, or *a priori* hypotheses of the researcher. On the contrary, the multivariate analysis allows the internal relationships of the database to emerge automatically. The term "multivariate" means, in fact, that these methods of analysis take simultaneously into account all the information and all the relationships. In this way, the analysis may respond to our questions, but also indicate the unexpected, if it exists. On the contrary, classical hypothesis testing statistics can only respond to the question: How different is an event in respect to a given probability distribution?

Exploring Genotoxicity Data

The importance of exploring the data without *a priori* hypotheses should be particularly emphasized. Let us consider the contribution of multivariate analyses to one of the problem that has occupied the mutagenists for years: the problem of how well the short-term tests are able to predict carcinogenicity.

In the first studies, Salmonella seemed to be capable of predicting the carcinogenicity of a high proportion of chemicals. Later on, more chemicals of different classes were studied, and this predictive ability considerably declined; consequently, the new problem of finding one or more short-term tests complementary to Salmonella for predicting carcinogenicity arose. For many years, tests complementary to Salmonella were sought among those with different genetic end point and phylogenetic position (e.g., chromosomal aberrations in mammalian cells). This search was largely based on the hypothesis that tests based on different genetic end points and different types of cells should respond differently to chemicals.

To test this hypothesis, we have performed a number of multivariate analyses of the most important genotoxicity data bases: International Program for the Evaluation of Short-Term Tests for Carcinogens (IPESTTC) (*3*), International Program for Chemical Safety (IPCS) (*4*), and the U.S. National Toxicology Program (NTP) (*5*). All our analyses (*6,7*) cogently showed that tests with different genetic end points and phylogenetic positions can respond in a similar way to the same set of chemicals, and vice versa: the difference in the performances of assays does not directly depend on differences in genetic end point or type of cells. This is particularly evident in the results of the NTP: here, the assay most similar to Salmonella (STY) is the chromosomal aberration test in CHO cells (CHA), which differs from Salmonella for both genetic end point and type of cells (bacterial instead of mammalian cells). The multivariate analyses of IPESTTC and IPCS pointed out the same result.

On the other hand, this disagreement between theories and experimental results does not automatically imply that there are no differences between the tests. For example, our analysis of IPESTTC data indicated the presence of three large families of short-term tests, different for their profiles of responses to the chemicals (*6,7*). We recall here this specific study because the IPESTTC is the only comparative program in which many different tests, both *in vitro* and *in vivo*, were studied simultaneously. The three groups of tests were *a*) a cluster including all *in vivo* assays, which gave positive responses for a limited number of chemicals; *b*) a cluster including Salmonella, together with many of the most widely used *in vitro* assays (e.g., chromosomal aberrations and sister chromatid exchange [SCE] in CHO cells, mouse lymphoma mutation, unscheduled DNA synthesis [UDS] in human fibroblasts). These tests showed positive responses for the chemicals positive in the *in vivo* short-term tests and were also sensitive to a number of other chemicals; *c*) a third cluster, including other *in vitro* assays (e.g., mutation in *S. cerevisiae* XV185-14C, Syrian hamster embryo cell transformation, *E. coli* polA), which were sensitive to the chemicals positive in the two other clusters of short-term tests but were also sensitive to other chemicals. The resulting view is that there are differences between the different tests, and these differences consist in a different sensitivity to a common, underlying property of chemicals, which can be called genotoxicity. The difference between tests is mostly quantitative in terms of how sensitive to genotoxins they are, qualitative differences in terms of types of genetic damage are only secondary. To correctly appreciate these differences, we should not use the traditional classification of tests according to genetic end point and phylogenetic position, which may be useful for other purposes, but we should shift to new categories, represented by the clusters of tests with homogeneous responses to the chemicals. These alternative categories automatically emerge from the experimental data, when studied with appropriate analysis methods.

Over the years, the idea that the type of genetic end point does not directly determine the performance of an assay and that assays based on chromosomal aberrations do not necessarily have a performance different than that of Salmonella has slowly taken shape [see for example, Ashby (*8*), or the recent proposal of the U.S. EPA (*9*) for new mutagenicity testing guidelines]. However, clear evidence was already provided by data available 10 years ago (e.g., by the IPESTTC data), but this evidence was not promptly perceived or accepted, thus slowing the progress of research. Probably, the present conclusions would have been reached more easily and quickly with the aid of the appropriate means for data analysis.

Modeling Genotoxicity Data

Multivariate methods can also be used for a different purpose: the mathematical modeling of data. We have exploited this ability of multivariate methods to study the specific problem of the comparison of different databases. Over the years, a large amount of information has been generated, both through individual studies and large, comparative studies. In the various comparative studies, often the same tests have been studied, but with different sets of chemicals. How can we compare the results of the different studies when the reference frame (that is, the chemicals) is different? How can we separate what is invariant from what is peculiar to the specific database?

We have considered the four assays studied in the NTP (STY, CHA, mouse lymphoma mutation [MLY], and SCEs in CHO cells) because they are also reported in the IPESTTC, IPCS, and Gene-Tox. The Gene-Tox data used here refer to a subset of chemicals reported in Palajda and Rosenkranz (*10*). The following is a description of our preliminary results (manuscript in preparation).

A simple way of comparing two assays based on the results of a set of chemicals, is to count the number of chemicals for which they give different results. The ratio of chemicals with different results to total number of chemicals is the Hamming distance between the two assays. If we compute the Hamming distance between all pairs of assays, then we obtain a Hamming distance matrix that completely summarizes the relationships between assays in a given database (*11*).

Even though the databases cannot be compared directly to each other because they are based on different sets of chemicals, with this approach we obtain distance matrices that are homogeneous, both formally and substantially. Rows and columns have the same meaning in each matrix and are therefore comparable. Each matrix defines a relationship pattern; if we compare these matrices to each other, we can see if the test relationships vary in the different databases.

A simple way of performing such a comparison is to calculate correlation coefficients between each pair of distance matrices. The resulting correlation-coefficient matrix gives the global similarities of the four databases (Table 1).

The matrix was studied by factor analysis, which gave a map of the similarities among databases (Fig. 1). IPESTTC is close (hence similar) to NTP, whereas IPCS and Gene-Tox express different relationships among tests. NTP and IPESTTC are based on sets of chemicals belonging to different chemical classes and are supposed to be samples of the universe of chemicals; in this way, they resemble each other. IPCS essentially consists of carcinogens selected because they are negative in Salmonella; thus, the IPCS is biased toward a specific goal and is not aimed at being representative of the universe of chemicals. In fact, in the map, IPCS is far from NTP and IPESTTC. This subset of Gene-Tox chemicals also includes many different chemical classes, like NTP and IPESTTC, but refers to chemicals assayed in a period in which the chemicals studied were selected mainly because of suspicions concerning their genetic activity or carcinogenicity. This bias is accounted for by the position of Gene-Tox in the factorial map. This result agrees well with what is known about the databases. This is important because it demonstrates exactly how sensitive this method of analysis is; hence, we can confidently use this approach in other situations in which we do not know much *a priori*. Moreover, it gives a precise, quantitative measure of the differences between databases; this is not possible with nonmathematical approaches.

Table 1. Correlation coefficients between databases.[a]

	NTP	IPESTTC	IPCS	Gene-Tox
NTP	1.000			
IPESTTC	0.628	1.000		
IPCS	-0.014	-0.112	1.000	
Gene-Tox	0.465	0.520	0.430	1.000

[a]NTP, U.S. National Toxicology Program; IPESTTC, International Program for Evaluation of Short-Term Tests for Carcinogens; IPCS, International Program for Chemical Safety.

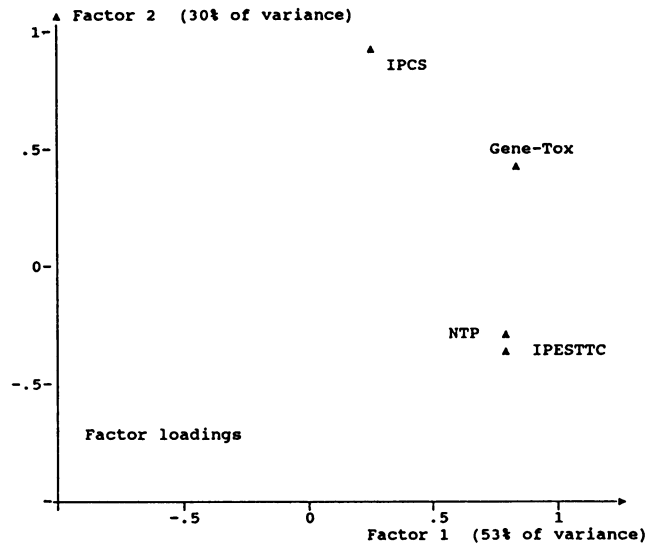

FIGURE 1. Relationships between databases.

After this global picture, we examined in more detail the problem of comparing different databases. We studied with separate factor analyses the four Hamming distance matrices that describe the test relationships in the four databases. The factors obtained summarized these relationships between tests: the number of factors was 1, 2, 2, and 2 for NTP, IPESTTC, IPCS, and Gene-Tox, respectively.

We compared these new variables (i.e., factors) to each other with a further factor analysis. The analysis indicated that all the information derived from the four databases can be summarized into two new factors. Figure 2 reports the position of the tests on factor 1, which describes the most important part of the information. It is evident that STY responds to the chemicals in a way similar to that of CHA, whereas MLY and SCE are similar to each other.

Because of the procedure used, this result of factor analysis can be considered as the best summary of the part of information that is invariant and repeated in the four databases. In other words, the similarities between tests shown by factor analysis are the result of a progressive search for the evidence common to all the databases. The importance of this result should be emphasized: an indication common to such a large amount of data is certainly the most reliable basis for any further investigation (aimed at elucidating biological mechanisms or at applications such as risk assessment, etc.).

Conclusions

In conclusion, the studies reported here show very clearly how biology and mathematical modeling can be combined and a true interdisciplinarity can be attained. Biology provides information

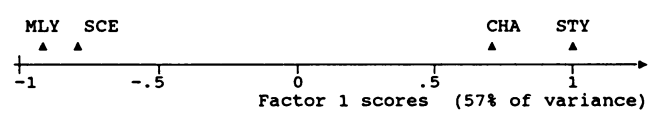

FIGURE 2. Overall relationships between assays (summary of four databases).

about the phenomena; mathematical modeling formalizes and organizes the information and precisely defines the relationships and points out the elements that play a role in the phenomenon. The advantages of the mathematical language should be strongly stressed: first, it has is the ability to describe small differences with higher flexibility and precision than with natural language, and second, it has the capability to manipulate and explore the selected features in an objective and flexible way.

REFERENCES

1. Lebart, L., Morineau, A., and Warwick, K. M. Multivariate descriptive statistical analysis. John Wiley and Sons, New York, 1984.
2. Benigni, R., and Giuliani, A. Multivariate analyses in genetic toxicology. In: Applied Multivariate Analysis in SAR and Environmental Studies (J. Devillers and W. Karcher, Eds.), Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991, pp. 347–376.
3. de Serres, F. J., and Ashby, J. Eds. Evaluation of Short-Term Tests for Carcinogens, Report of the International Collaborative Program. Progress in Mutation Research, Vol. 1. Elsevier/North-Holland, Asterdam, 1981.
4. Ashby, J., de Serres, F. J., Draper, M., Ishidate, M., Margolin, B. H., Matter, B. E., and Shelby, M. D. Evaluation of Short-Term Tests for Carcinogens, Report of the International Program on Chemical Safety Collaborative Study on In Vitro Assays. Progress in Mutation Research, Vol. 5. Elsevier/North-Holland, Amsterdam, 1985.
5. Tennant, R. W., Margolin, B. H., Shelby, M. D., Zeiger, E., Haseman, J. K., Spalding, J. W., Caspary, W., Resnick, M., Stasiewicz, S., Anderson, B., and Minor, R. Prediction of chemical carcinogenicity in rodents from in vitro genetic toxicity assays. Science 236: 933–941 (1987).
6. Benigni, R., and Giuliani, A. Predicting carcinogenicity with short-term tests: biological models and operational approaches. Mutat. Res. 205: 227–236 (1988).
7. Benigni, R., and Giuliani, A. Statistical exploration of four major genotoxicity data bases: an overview. Environ. Mol. Mutagen. 12: 75–83 (1988).
8. Ashby, J. The prospects for a simplified and internationally harmonized approach to the detection of possible human carcinogens and mutagens. Mutagenesis 1: 3–16 (1986).
9. Dearfield, K. L., Auletta, A. E., Cimino, M. C., and Moore, A. M. Considerations in the U.S. Environmental Protection Agency's testing approach for mutagenicity. Mutat. Res., in press.
10. Palajda, M., and Rosenkranz, H. S. Assembly and preliminary analysis of a genotoxicity data base for predicting carcinogens. Mutat. Res. 153: 79–134 (1985).
11. Benigni, R. The National Toxicology Program data on the in vitro genetic toxicity tests using multivariate statistical methods. Mutagenesis 4: 412–419 (1989).