Discriminant Function Analyses of Liver-Specific Carcinogens

Richard D. Beger,*,[†] John F. Young,[‡] and Hong Fang[§]

Division of Chemistry and Division of Biometry & Risk Assessment, Food & Drug Administration,
National Center for Toxicological Research, Jefferson, Arkansas 72079,
and Logicon ROW Sciences, Jefferson, Arkansas 72079

Received December 5, 2003

The ability to predict organ-specific carcinogenicity would aid FDA reviewers in evaluating new chemical applications. A NCTR liver cancer database (NCTRlcdb) containing 999 compounds has been developed with three sets of descriptors. The NCTRlcdb has Cerius2, Molconn-Z, and ^{13}C NMR descriptors for each compound. Each compound in the database was assigned a liver cancer or a nonliver cancer classification. Compounds within the NCTRlcdb were evaluated for liver-specific carcinogenicity using partial least squares principal component discriminant function (PLS-DF) modeling. PLS-DF models based on estimated a priori classification probabilities of 0.29 for liver cancer and 0.71 for noncancer yielded an overall predictability of 70.6% which was comprised of a liver cancer sensitivity of 18.8% and a noncancer specificity of 90.8%. PLS-DF models based on equal a priori classification probabilities, 0.50 for liver cancer and 0.5 for noncancer, yielded an overall predictability of 61.0% which was comprised of a liver cancer sensitivity of 50.5% and a noncancer specificity of 65.3%.

INTRODUCTION

Time and cost considerations do not make it feasible to carry out carcinogenic bioassays on every molecule. However, using the information from cancer bioassays and the ability to build structure−activity relationships, an untested molecule might be evaluated, reducing the burden for all encompassing testing. A NCTR liver cancer database (NCTRlcdb) of cancer-tested molecules has been developed[1,2] from a subset of the carcinogenic potency database (CPDB) of Gold and colleagues.[3] The NCTRlcdb was developed in order to develop structure−activity relationship models of liver cancinogens. In the NCTRlcdb each chemical was assigned a toxicity classification (liver cancer = 1, nonliver cancer = 0). For structure−activity relationship (SAR) models, the chemical structure is needed in molecular format for subsequent generation of chemical descriptors. Developers of the NCTRlcdb compiled the information from the CPDB through the use of a hierarchical scheme to reduce multiple records representing each gender, species, route of administration, and organ-specific toxicity into a single record for each study.[1] The development of the NCTRlcdb included a structural cleanup that involved removal of all molecules that would interfere with the generation of SAR type descriptors from commercial software programs; i.e., inorganic compounds, mixtures, organometallics, and counterions.[1]

Partial least squares discriminant function (PLS-DF) modeling has been used previously to determine which variables discriminate between two or more groups.[4,5] This type of modeling is especially important when there are many variables and a limited number of groups. PLS-DF models can be built by using a priori classification probabilities set to an estimated probability based on the distribution of training set cases or by using a priori classification probability set to equal levels for each group. The NCTRlcdb had an unequal distribution of liver cancer and nonliver cancer causing molecules, which can cause problems in the training of a model. In practice, the unequal number training cases in each group can be a reflection of the true distribution of the population of cases or a random result of the sampling procedure. If the number of training cases in each group was a true distribution, an estimated a priori classification probability should be used. If the number of training cases in each group was not a true distribution of cases and the true distribution of cases are unknown, an equal a priori classification probability should be used. If the number of training cases in each group was not a true distribution of cases but the true distribution was known, a priori classification probability should be set to the true distribution. In the case of NCTRlcdb, a true distribution of liver cancer causing molecules was unknown, and the data set was biased to compounds believed to be noncancerous. In this paper the PLS-DF technique with an equal a priori classification probability was able to produce predictive models of liver cancer.

METHODS

Structures were available for all 999 of the NCTRlcdb molecules. Structural descriptors were generated from software obtained from Cerius2 (Accelrys Inc., San Diego, CA) and Molconn-Z (eduSoft, LC, Ashland, VA). Each of the descriptor sets were used independently in separate PLS-DF models. In addition each molecule had its corresponding ^{13}C NMR spectra predicted using ACD Labs CNMR Version 5.0 software (ACD/Labs, Toronto, Canada). The predicted

* Corresponding author phone: (870)543-7080; fax: (870)543-7686; e-mail: rbeger@nctr.fda.gov. Corresponding author address: Division of Chemistry, Food & Drug Administration, National Center for Toxicological Research, Jefferson, AR 72079.

[†] Division of Chemistry, Food & Drug Administration.

[‡] Division of Biometry & Risk Assessment, Food & Drug Administration.

[§] Logicon ROW Sciences.

10.1021/ci0342829 Published 2004 by the American Chemical Society
Published on Web 03/31/2004

NMR spectra were calculated by a substructure similarity technique called HOSE,[6] which correlates similar structures with similar NMR chemical shifts. Therefore, the errors produced in the simulated NMR spectra were propagated through to the similar structures found in the training set of the spectrometric data−activity relationship (SDAR) models. This conveniently reduced the effective error when using the training set to predict unknown sample affinities for compound spectra predicted using the same HOSE routine.

Unassigned 1D ^{13}C NMR chemical shifts were segregated into bins over a 0−240 ppm range. The ^{13}C NMR spectra were saved as a set of ordered pairs: chemical shift frequencies in ppm and peak areas. The area under a specific chemical shift frequency was first normalized to an integer value. A nondegenerate frequency was assigned a value of 100, a doubly degenerate frequency (2 ^{13}C NMR chemical shifts at the same frequency) was assigned 200, and so forth. The number 100 was selected arbitrarily in order to normalize the peak intensity in ^{13}C NMR spectroscopic data. This initial normalization was done so that (1) all the spectra would have a similar signal-to-noise ratio and (2) to eliminate line width variations due to differences in NMR instrumental field strengths, shimming, coupling to protons, temperature, pH, or solvent. The bin defined the number of significant and distinct chemical shift peaks with a ppm range. Previous SDAR models have shown that a 1 ppm bin size was very functional.[4,5] Using a 1 ppm spectral width for the bins, 223 out of the 240 spectral bins had nonzero ^{13}C chemical shifts intensities for at least one chemical.

Partial least squares principal component discriminant function (PLS-DF) was applied to each set of descriptors in the NCTRlcdb.[4,5] Because there are so many descriptors and only liver cancer and nonliver cancer groups, PLS-DF modeling can be subject to overfitting and is especially important to validate through cross-validation techniques and external tests. Therefore, each model was built based upon the training set (75% of the database) of chemicals and then applied to the corresponding test set (remaining 25% of the database). This was done four times with four different test sets, so each chemical in the database was in a test set only once. Each test set had approximately the same numbers of liver cancers (72−73) and nonliver cancer chemicals (177−178). The PLS-DF liver cancer models were trained using three of the test sets and used the fourth test set as an independent external test set.

The descriptors are Fisher-weighted prior to pattern recognition to emphasize those descriptors that are most important for differentiating the endpoint classes. Fisher-weighting (FW) for a descriptor was defined as

$$\text{FW (descriptor)} = \frac{\sum_{i}^{g\text{-}1} \sum_{j=i}^{g} (\langle X_i \rangle - \langle X_j \rangle)^2 / (\sigma_i^{\,2} + \sigma_j^{\,2})}{g\ ((g\text{-}1)/2)}$$

where g was the number of classification categories, $\langle X_i \rangle$ was the average intensity for category i, and σ_i was the variance about the mean of category i. For each descriptor, the variance between categories was divided by the variance within categories. The resulting dividend became a weighting factor that had a magnitude larger than one when a particular descriptor has a role in distinguishing groups. Fisher weighting of all of the descriptors before pattern recognition increased the power of discriminant analysis for classification purposes. This weighting was particularly important in this application because it effectively de-emphasized irrelevant descriptor information. Each descriptor was multiplied by its weighting factor to yield descriptors that were more sensitive to subtle but significant variations. After the descriptors were multiplied by their Fisher weight, principal component (PC) extraction by correlation analysis was completed using Statistica version 6 (Statsoft, Tulsa, OK). Results of the PLS-DF models using principal components based on variance analysis were not reported but gave predictive results that were generally within 3% of the PLS-DF correlation model. The traditional PLS-DF analysis models were then built using Statistica software. Cerius 2 PLS-DF models were built using the first 80 PCs which contained 93.6% of the total variation in the Cerius 2 data set. CNMR set PLS-DF models were built selecting from the first 80 PCs which contained 98.3% of the variation in the CNMR data set. CNMR PLS-DF models did not forward select PCs after the 70th PC. Molconn-Z PLS-DF models were built using the first 100 PCs which contained 93.5% of total variation in the Molconn-Z data set. The PLS-DF models were built using forward regression on the PCs, and a PC was added to the model if it increased the F-score of the model by 1.0. Typically, 30−40 PCs were forward selected under these conditions to produce the trained PLS-DF model. A PLS-DF model was built for each training set and then cross-validated using the associated external test set. The process of building a PLS-DF training set and cross-validating with an external test set was done four times.

For each descriptor set, two PLS-DF models were built; one model using a priori classification probabilities set to an estimated (based on the NCTRlcdb population fraction of 0.29 for liver cancer and 0.71 for noncancer) and a second model using a priori classification probability set to equal levels (0.50 for liver cancer and 0.5 for noncancer). Models were built using a priori classification of estimated and equal because in the case of liver cancer prediction, a true distribution was unknown. The relationship between a priori and case weights are quite complex[7] and can produce models with widely varying classification results. Therefore, these PLS-DF models need to be tested rigorously.

RESULTS

Figure 1A is a three-dimensional display for the first three principal components of 999 chemicals using the Cerius 2 descriptor set with no Fisher weighting. Figure 1B is a three-dimensional display for the first three principal components of 999 chemicals using the Cerius 2 descriptor set with Fisher weighting. Whether or not Fisher weighting was used, the first principal component has 23.1% of the Cerius 2 descriptor variation. The black points are for noncancerous compounds, and the gray points display the compounds that cause liver cancer. Figure 1A,B shows that the compounds used in these models of liver cancer are quite diverse. Figure 1A shows a lot of overlap between cancer and noncancer causing compounds. Figure 1B shows that the first Fisher weighted principal component attempts to split the compounds into two subgroups; still there is a lot of overlap between cancer and noncancer causing compound in each subgroup.

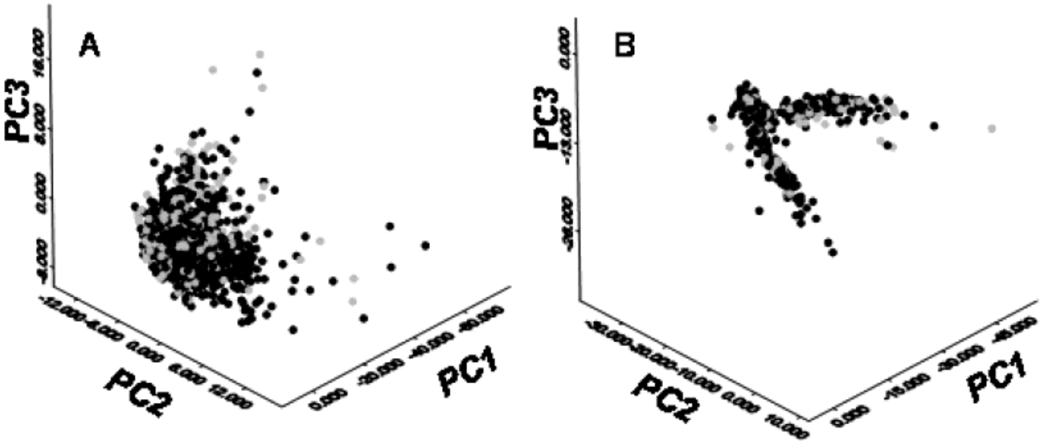

Figure 1. A: Graphical display of the first three principal components of 999 chemicals using the Cerius 2 descriptor set with no Fisher weighting. 1B: Graphical display of the first three principal components of 999 chemicals using the Cerius 2 descriptor set with Fisher weighting. The black points are for noncancerous compounds, and the gray points display the compounds that cause liver cancer.

Table 1. Results of the PLS-DF Predictions Made for the 4 Training Sets

PLS-DF method	descriptors	sensitivity % liver cancer correct	specificity % noncancer correct	% total correct
estimated	Cerius2	33.2 ± 3.1	93.5 ± 1.2	76.0 ± 1.1
estimated	Molconn-Z	29.1 ± 5.7	94.2 ± 1.1	75.3 ± 1.2
estimated	CNMR	15.9 ± 6.7	96.5 ± 0.5	73.7 ± 2.0
equal	Cerius2	66.8 ± 2.7	74.0 ± 1.9	71.9 ± 1.4
equal	Molconn-Z	59.0 ± 5.8	64.0 ± 0.9	62.6 ± 1.2
equal	CNMR	53.5 ± 2.6	67.3 ± 1.0	66.3 ± 0.8

Table 2. Results of the PLS-DF Predictions Made for the 4 External Test Sets

PLS-DF method	descriptors	sensitivity % liver cancer correct	specificity % noncancer correct	% total correct
estimated	Cerius2	25.9 ± 5.9	90.0 ± 2.9	71.8 ± 8.3
estimated	Molconn-Z	24.5 ± 4.1	87.0 ± 4.3	68.9 ± 2.8
estimated	CNMR	5.9 ± 2.1	95.4 ± 2.4	69.5 ± 2.2
equal	Cerius2	56.6 ± 4.9	68.9 ± 3.3	65.3 ± 1.7
equal	Molconn-Z	59.0 ± 5.8	64.0 ± 0.9	62.6 ± 1.2
equal	CNMR	35.9 ± 5.4	63.1 ± 2.2	55.2 ± 1.1

Table 3. Results of the PLS-DF Model Predictions Made for the 4 External Test Sets with Three Randomized Liver Cancer Training Sets

PLS-DF method	descriptors	sensitivity % liver cancer correct	specificity % noncancer correct	% total correct
estimated	Cerius2	19.9 ± 18.2	81.5 ± 13.1	63.7 ± 4.8
equal	Cerius2	52.2 ± 24.4	51.3 ± 23.8	51.5 ± 10.3

The PLS-DF analysis scheme was applied to the three Fisher weighted descriptor sets (i.e., Cerius2, Molconn-Z, and CNMR). The results of each set of descriptors are presented in Table 1 as a mean and standard deviation for the four training sets of compounds for both estimated and equal a priori classification probability models. The total percent correct in the training set ranged from 74 to 76% for the estimated a priori models and 63−72% for the equal a priori models. All trained PLS-DF estimated a priori classification models had higher overall accuracies and noncancer accuracies than the corresponding trained PLS-DF equal a priori classification models.

The results for the four test sets of compounds from each set of descriptors are presented in Table 2 as a mean and standard deviation for both estimated and equal a priori classification probability models. The total percent correct in the external test set ranged from 55 to 72% among the various analysis schemes and descriptors. All external test set predictions were better at specificity (predicting noncancers correctly) than sensitivity (predicting liver cancers correctly). All PLS-DF external predictions made using estimated a priori classification models had higher overall accuracies and noncancer accuracies than the corresponding external predictions made using equal a priori classification models. All three PLS-DF equal classification models had higher overall liver cancer accuracies than the corresponding PLS-DF estimated a priori classification models. PLS-DF models based on equal a priori classification probabilities yielded an overall predictability of ∼61% which was comprised of a liver cancer sensitivity of ∼51% and a nonliver cancer specificity of ∼65%. PLS-DF models based on estimated a priori classification probabilities yielded an overall predictability of ∼71% which was comprised of a liver cancer sensitivity of ∼19% and a nonliver cancer specificity of ∼91%. Cerius2 descriptors tended to give the most accurate test result predictions, followed by Molconn-Z descriptors.

PLS-DF models trained using Cerius 2 descriptors were subject to randomized liver cancer data sets to determine the effects of randomized training data. The PLS-DF models were trained using the 75% of the randomized liver cancer data, and then these models were used to predict the remaining 25% of the chemicals using the real liver cancer chemical status. Each randomization set was applied to the same four training chemical sets as used earlier in training with 100% liver cancer chemical status. This randomization process was done three separate times with three orthogonal randomized training sets. The randomization results of the Cerius 2 descriptors are presented in Table 3 as a mean and standard deviation for the four external test sets of compounds for both estimated and equal a priori classification probability models. The total percent correct in the training set ranged from 64% for the estimated a priori models to 52% for the equal a priori models. Randomized liver cancer data caused the Cerius 2 estimated a priori model overall accuracy to drop by 8.1%, the sensitivity to drop by 6.0%, and the specificity to drop by 8.5%. Randomization caused the Cerius 2 equal a priori model overall accuracy to drop by 13.8%, the sensitivity to drop by 4.4%, and the specificity to drop by 17.6%. Of course making a nonliver cancer prediction for all chemicals would result in 71% overall accuracy, 0% sensitivity, and 100% specificity would be better than any of the randomization models, but an all noncancer prediction gives no insight into what chemical characteristics cause liver cancer.

DISCUSSIONS

The PLS-DF model results based on Cerius2, Molconn-Z, and CNMR data were very similar. The higher a priori classification probability bias is for liver cancer, the more likely it is to predict liver cancer correctly as is seen in all three equal biased PLS-DF models. The same rationale can be applied to describe the classification of nonliver cancers, in all three estimated PLS-DF models. The interesting phenomenon in these PLS-DF models is changing the liver cancer bias from 0.29 to 0.50 increases the accuracy of cancer prediction by 30−32% for Cerius2, Molconn-Z, and CNMR data, while the overall accuracy of the PLS-DF models only falls by 6−14%. The choice of the PLS-DF model will depend on what information is desired. If total or nonliver cancer prediction accuracy is the most important part of the model, then the estimated bias PLS-DF models are better. If cancer prediction is the most important part of a model, as it would be for FDA reviewers, then the equal bias PLS-DF models are better. Since the PLS-DF models with an equal a priori classification probability bias for liver cancer have predictabilities greater than 50% for both liver cancer and nonliver cancer, they appear to be learning what types of molecular characteristics result in liver cancer. The estimated models that tend to predict mostly noncancer chemicals were not affected by randomization as much, whereas the equal models that seemed to learn more about the properties of cancer causing compounds were affected by the randomization a little more. The randomization results seem to demonstrate that the equal a priori models are better models of liver cancer causing compounds.

The overall predictability for the models especially for liver cancer sensitivity was not as accurate as hoped. The hope had been for a somewhat higher value. The Molconn-Z PLS-DF equal model had a predictive sensitivity for liver cancer of 59.0% which is not great but not bad. The PLS-DF models are comparing favorably to Splines (MARS) (Salford Systems, San Diego, CA), Rough Sets,[8] and Support Vector Machines[9] that were reported by Young et al.[1] The MARS liver cancer models based on nonlinear relationships yielded an overall predictability of ∼67% which was comprised of a liver cancer sensitivity of ∼32% and a specificity of ∼82%. The Rough Sets liver cancer analyses yielded an overall predictability of ∼57% which was comprised of a liver cancer sensitivity of ∼18% and a specificity of ∼74%. The support vector machine liver cancer analyses yielded an overall predictability of ∼66% which was comprised of a liver cancer sensitivity of ∼23% and a specificity of ∼84%.

1D ^{13}C NMR spectral data only contains local environment quantum mechanical information with no structural information, so it comes as no surprise that the diversity of molecules in this data set and the diversity in this liver cancer endpoint showed that 1D ^{13}C NMR descriptors were not good descriptors for liver cancer. The addition of ^{15}N NMR chemical shift information and the combination of structural information with NMR spectral information may improve NMR PLS-DF models of liver cancer.[10,11]

The overall predictability of the a priori estimated PLS-DF models using a priori estimated classification scheme was not especially satisfying, but the PLS-DF models using a priori equal classification scheme held promise. One reason that the results of modeling liver cancer were poorer than expected could be due to varying mechanisms of action in regard to tumor formation. Another reason might be due to the wide variety of chemical structures. Figure 1B showed that Fisher weighting split the compounds into two groups, but there is still a lot of overlap between chemicals that cause cancer and chemicals that do not cause liver cancer. This should mean that nonlinear modeling techniques would be important to modeling liver cancer. Although, the nonlinear MARS technique was superior to many linear modeling attempts of chemicals that cause liver cancer, the overall accuracy of MARS was very similar to all other models of liver cancer.[1] The poor overall accuracy performance by both linear and nonlinear modeling techniques may be due to the fact that many chemicals are metabolized to the active tumor causing molecule. We were hoping to find chemical properties within the three descriptor sets that would signify a cancer causing chemical, but the best PLS-DF models were only able to predict liver cancer compounds 60% of the time.

REFERENCES AND NOTES

(1) Young, J. F.; Tong, W.; Fang, H.; Xie, Q.; Pearce, B.; Hashemi, R.; Beger, R. D.; Cheeseman, M. A.; Chen, J. J.; Chang, Y. I.; Kodell, R. L. Building an organ-specific carcinogenic database for SAR analyses. *J. Toxicol. Environ. Health, Part A* **2004**, in press.

(2) Young, J. F.; Tong, W.; Fang, H.; Beger, R. D.; Chen, J. J.; Cheeseman, M. A.; Kodell, R. L. Computational predictive system for rodent organ-specific carcinogenicity. *Comput. Intell.: Methods Appl.* **2001**, CIMA'2001, 565−569.

(3) Gold, L. S.; Sawyer C. B.; Magaw, R.; Backman, G. M.; de Veciana, M.; Levinson, R.; Hooper, N. K.; Havender, W. R.; Bernstein, L.; Peto, R.; Pike, M. C.; Ames. B. N. A carcinogenic potency database of the standardized results of animal bioassays. *Environ. Health Persp.* **1984**, *58*, 9−319.

(4) Beger, R. D.; Freeman, J. P.; Lay, J. O., Jr.; Wilkes, J. G.; Miller, D. W. ^{13}C NMR and EI Mass spectrometric data to produce a predictive model of estrogen receptor binding. *Toxicol. Applied Pharmacol.* **2000**, *169*, 17−25.

(5) Shade, L.; Beger, R. D.; Wilkes, J. G. New computerized method for modeling binding affinities to the aryl hydrocarbon receptor using ^{13}C NMR spectra. *Environ. Toxicol. Chem.* **2003**, *22*, 501−509.

(6) Bremser, W. HOSE − a novel substructure code. *Anal. Chim. Acta* **1978**, *103*, 355−365.

(7) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees;* Wadsworth & Brooks/Cole Advanced Books & Software: Monterey, CA, 1984.

(8) Hashemi, R.; Pearce, B.; Arani, R.; Hinson, W.; Paule, M. A fusion of rough sets, modified rough sets, and genetic algorithms for hybrid diagnostic systems. In *Rough Sets and Data Mining: Analysis of Imprecise Data*; Lin, T. Y., Cercone, N., Eds.; Kluwer Academic Publishers: New York, 1997; pp 149−176.

(9) Vapnik, V. *Statistical Learning Theory*; John Wiley & Sons: New York, 1998.

(10) Beger, R. D.; Buzatu, D.; Wilkes, J. G.; Lay, J. O., Jr. Developing comparative structural connectivity spectra analysis (CoSCSA) models of steroid binding to the corticosteroid binding globulin. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1123−1131.

(11) Beger, R. D.; Buzatu, D.; Wilkes, J. G. Combining NMR spectral and structural data to form models of polychlorinated dibenzodioxins, dibenzofurans, and biphenyls binding to the AhR. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 727−740.

CI0342829