PLS Analysis of Distance Matrices To Detect Nonlinear Relationships between Biological Potency and Molecular Properties

Yvonne C. Martin,*,† C. Thomas Lin,*,‡ Chandanie Hetti,§ and Jerry DeLazzer⊥

Pharmaceutical Products Division, Abbott Laboratories, 100 Abbott Park Road, Abbott Park, Illinois 60064-3500

Received March 17, 1995®

Although the statistical method of partial least squares (PLS) is widely used for the analysis of the relationship between molecular properties and biological potency, it is recognized that PLS detects only linear relationships. We tested two types of properties: simulated univariate data and electrostatic molecular field as a function of Hammett σ constants. In both cases we compared relationships in which the function is linear, asymptotic, or rises to an optimum and then falls. We found that PLS analysis of the matrix of the distances between every pair of compounds detects all three types of relationships with the same quality of cross-validation. The successful application of the method requires that the distance matrices be constructed such that each contains only information about one property (for example, the electrostatic field around the functional group of interest). Carbo and Hodgkin similarities perform less well than distances.

Background

PLS, partial least squares, is a popular statistical method that is used to find relationships between a dependent property such as bioactivity and a large number of potential predictor properties.[1] Its strength is that through cross-validation it detects valid relationships but does not over-fit the relationship.[2] PLS is used in CoMFA, comparative molecular field analysis, a commonly used method for analyzing the relationships between biological potency and the 3D features of the molecules.[3,4] However, PLS has an important weakness in that it does not detect nonlinear relationships between the dependent property and potential predictors.

When the biological potency of a set of molecules depends on their pK_a's, the relationship will not always be linear, but may instead be parabolic or asymptotic.[5] Therefore, although CoMFA electrostatic fields can explain linear relationships with pK_a and forecast pK_a's of compounds not included in data sets,[6,7] we have been concerned that PLS and hence CoMFA would miss nonlinear relationships. Hence we were intrigued by a brief report that a PLS analysis of a matrix of the distances between every compound of a data set detected the nonlinear relationship between a physical property and biological potency.[8] Others use PLS to analyze similarity matrices and frequently find that the models are slightly statistically superior to PLS analysis based on the physical property data from which the similarity matrices were calculated.[9,10]

This report documents that nonlinear relationships between physical properties and biological potency can be recognized by PLS when (1) the physical properties are used to calculate a distance between each pair of compounds in the dataset and (2) the resulting distance matrix is analyzed with PLS. We will show that in order to detect the relationship, the distance matrices should be calculated such that properties that are intrinsically independent are not mixed into one distance measure.

As already recognized by others, the calculation of a distance matrix is very fast.[11] Hence the proposed method of analysis is much faster than tradiational PLS on the individual energy values of CoMFA fields.

Methods

The first dataset we analyzed contains simulated univariate data sets of 31 points in which the relationship between the dependent and independent properties is linear, asymptotic, or rises to an optimum and then falls, Figure 1. We deliberately constructed these data sets to include many approximate duplicates so that the power of cross-validation[12-15] is realized. These data are quite well fit by standard regression analysis using linear and quadratic functions of the independent variable: For the asymptotic case, the fitted R^2 is 0.99 and the root mean square error is 0.097, whereas the fit to a simple linear function gives statistics of 0.81 and 0.384. For the optimum case, the corresponding statistics are 0.93 and 0.200 for the nonlinear fit and 0.00 and 0.77 for the linear fit.

The Euclidean distance between two compounds **a** and **b** is given by eq 1:

$$D_{ab} = \sqrt{\sum_k (P_{a,k} - P_{b,k})^2} \quad (1)$$

where there are n properties P_k. In univariate data sets, this distance reduces to the absolute value of the difference between their properties, eq 2:

$$D_{ab} = |P_a - P_b| \quad (2)$$

The Hodgkin–Richards similarity metric S_{ab} is defined as[9]

$$H_{ab} = \frac{2\sum_k P_{a,k}P_{b,k}}{\sum_k (P_{a,k})^2 + \sum_k (P_{b,k})^2} \quad (3)$$

H_{ab} for the univariate case is thus calculated by eq 4:

$$H_{ab} = \frac{2P_aP_b}{P_a^2 + P_b^2} \quad (4)$$

Since H_{ab} is indeterminant if P_a and P_b are both zero, we used mean-centered P values for the calculations in Table 2.

Similarity or distance matrices are often calculated such that all properties are combined into one measure. This may

† Computer Assisted Molecular Design Project, D-47E, AP10.
‡ Biostatistics Department, D-431, AP9A.
§ Department of Mathematics, Statistics, and Computer Science, Northern Illinois University, Dekalb, IL.
⊥ Computational Molecular and Cell Sciences Department, D-42T, AP6B.
® Abstract published in *Advance ACS Abstracts*, July 1, 1995.

0022-2623/95/1838-3009$09.00/0

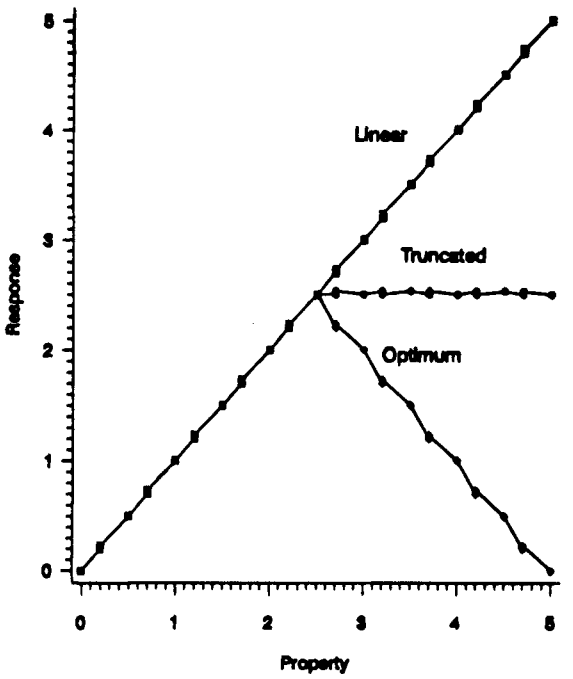

Figure 1. The simulated data used to test the ability of PLS to detect nonlinear relationships. The symbols show the data; the lines connect adjacent points.

add irrelevant data to the measure and perhaps obscure the signal. To test the effect of adding irrelevant data, we added to the data sets in Figure 1 a column of randomly scrambled data that is not correlated to the independent property. We analyzed the combined data using both (1) a distance matrix calculated from the two properties and (2) two independent distance matrices, one calculated from each property.

To test the ability of PLS to detect nonlinear relationships between a dependent variable and electrostatic fields, we used the 49 benzoic acids previously reported to show a good fit with Hammett σ.[6] These structures had been optimized with AM1 which is also the source of the partial atomic charges. The fields were calculated in Sybyl Version 6.04a (Tripos, Inc., St. Louis, MO); both steric and electrostatic fields were truncated at 30 kcal/mol; the exponent of the steric repulsion term was 12; a distance-dependent dielectric was used; the electrostatic terms were dropped in regions of high steric field; and a 2 Å spacing was used except for the analyses reported in Table 6 where a 1 Å spacing was used. The lattice extended beyond the molecules by *ca.* 4 Å.

Since the original work was done using fields calculated by Grid[16] but we now use Sybyl CoMFA, we reanalyzed the benzoic acid data using the default region calculated by Sybyl and a 1 Å spacing. The resulting CoMFA model (PLS model 37, Table 6) had six significant latent variables, a cross-validated standard error and R^2 of 0.126 and 0.89, respectively, and fitted values of 0.051 and 0.98. This model compares to the literature analysis with seven latent variables and a fitted R^2 of 0.98.[6]

We investigated fits to σ, to σ truncated at 0.20, and to σ with an optimum at 0.20 with the same upward and downward slopes. Table 1 shows the data used. All PLS analyses were conducted with SAMPLS (Quantum Chemistry Program Exchange, Bloomington, IN)[11] as interfaced to Sybyl. Since part of the SAMPLS algorithm is to calculate a Euclidean distance between every pair of observations, we imported this distance matrix into a Sybyl spreadsheet and used it for the distance-based PLS calculations. To calculate the covariance matrix, we used eq 5:[11]

$$C_{ab} = \sum_k (P_{a,k} - \overline{P_k})(P_{b,k} - \overline{P_k}) \quad (5)$$

where $\overline{P_k}$ is the mean value of the kth property.

The covariance can be calculated from the distance matrix D_{ab} as follows:

$$C_{ab} = -{}^1/_2(D_{ab}{}^2 - D_{a\cdot}{}^2 - D_{\cdot b}{}^2 + D_{\cdot\cdot}{}^2) \quad (6)$$

where $D_{a\cdot}{}^2$ and $D_{\cdot b}{}^2$ denote the averages of the squared distances over row a and column b, respectively, and $D_{\cdot\cdot}{}^2$ an average over all rows and columns.

We calculated the Hodgkin and Carbo similarities using ASP 3.02 (Oxford Molecular Ltd., Sandford-on-Thames, Oxford, U.K.) based on a 1.0 Å grid that extends 4.0 Å beyond the molecules. Since this version of ASP does not allow region selection, we also used Sybyl to calculate the fields at a 1 Å spacing for comparison.

We report the statistics for the models with the lowest leave-one-out cross-validated standard error, provided that adding the last latent variable to the model decreases the cross-validated standard error by at least 5% compared to the previous low. This criterion is more conservative than the Sybyl default, but it ensures that each component increases the cross-validation substantially.

The cluster analyses were performed using SAS PROC CLUSTER (SAS Institute Inc., Cary, NC), complete linkage algorithm, without normalization of the input scores. The optimal number of clusters was selected based on the local optimum of the pseudo F value.

Results

In Table 2 we show a comparison between PLS using the original variable and PLS based on the covariances, similarities, or distances between the compounds in property space. Models 5 and 9, relating to PLS on the original property, illustrate that the more nonlinear the data, the poorer the statistics using PLS. In contrast, distance-based PLS, models 8 and 12, performs as well on nonlinear data as it does on linear data. Recall that the fitted RMSE for the ordinary regression analysis quadratic fit is 0.097 for the asymptotic relationship and 0.200 for the optimum relationship. The corresponding standard errors are 0.029 and 0.053 for the distance-baed PLS, a clearly superior fit. Table 2 also shows that, for this univariate example, PLS based on the covariance of Hodgkin similarity matrices did not model the nonlinear data, PLS models 6, 7, 10, and 11.

Consideration of the cross-validated standard errors in Table 3 suggests that the improvement produced by using distance matrices is dependent on the construction of the distance matrix. If irrelevant properties are included, PLS models 13, 15, and 17, the power of distance-based PLS to detect the underlying relationships is reduced compared to that of the corresponding PLS models 4, 8, and 12. Indeed, even including extra columns of random numbers decreases the precision of the results, PLS models 14, 16, and 18.

From Table 4 we see that nonlinear relationships with electrostatic fields surrounding the carboxyl group are also detected by distance-based PLS, models 24 and 27. For the nonlinear cases, the cross-validated standard error is halved for distance-based PLS compared to PLS based on electrostatic fields, PLS models 24 *vs* 22 and 27 *vs* 25. This table also shows that PLS analysis of the covariance matrix does not produce superior results compared to PLS analysis of the original property matrix.

In Figure 2 we show a plot of the Hammett σ constant as a function of the Hammett σ forecast from the distance-based PLS analysis. Although the cross-validation is quite good, note the scatter of the points. This reflects the fact that electrostatic fields calculated

Table 1. Benzoic Acid Data Used for the Simulations and the Cluster Membership Based on Different Molecular Descriptors

	property			cluster basis			
substituent	Hammett σ	truncate at $\sigma = 0.20$	optimum at $\sigma = 0.20$	distance	similarities	score PLS model 24	score PLS model 27
p-NH_2	−0.66	−0.66	−0.66	6	1	6	6
p-OH	−0.37	−0.37	−0.37	6	1	6	6
p-OMe	−0.27	−0.27	−0.27	6	1	6	6
p-*t*-Bu	−0.20	−0.20	−0.20	1	1	1	1
p-CH_3	−0.17	−0.17	−0.17	1	1	1	1
m-NH_2	−0.16	−0.16	−0.16	1	1	1	1
m-*t*-Bu	−0.15	−0.10	−0.15	1	1	1	1
p-Et	−0.15	−0.15	−0.15	1	1	1	1
m-CH_3	−0.07	−0.07	−0.07	1	1	1	1
me-Et	−0.07	−0.07	−0.07	1	1	1	1
H	0.00	0.00	0.00	1	1	1	1
p-SMe	0.00	0.00	0.00	1	1	1	1
p-F	0.06	0.06	0.06	1	4	1	2
m-CH_2I	0.10	0.10	0.10	1	1	1	1
m-CH_2Cl	0.11	0.11	0.11	1	1	1	1
p-CH_2I	0.11	0.11	0.11	1	1	1	1
m-CH_2Br	0.12	0.12	0.12	1	1	1	1
m-OH	0.12	0.12	0.12	1	1	1	1
m-OMe	0.12	0.12	0.12	1	1	1	1
p-CH_2Cl	0.12	0.12	0.12	1	1	1	1
p-CH_2Br	0.14	0.14	0.14	1	1	1	1
m-SMe	0.15	0.15	0.15	1	1	1	1
p-SH	0.15	0.15	0.15	1	1	1	1
p-I	0.18	0.18	0.18	1	1	1	1
p-Br	0.23	0.20	0.17	1	1	1	1
p-Cl	0.23	0.20	0.17	1	1	1	1
m-SH	0.25	0.20	0.15	1	1	1	1
m-F	0.34	0.20	0.06	1	2	1	1
m-I	0.35	0.20	0.05	1	1	1	2
p-OCF_3	0.35	0.20	0.05	6	4	3	4
m-Cl	0.37	0.20	0.03	1	1	1	1
m-OCF_3	0.38	0.20	0.02	5	2	5	2
m-Br	0.39	0.20	0.01	1	1	1	2
m-SCF_3	0.40	0.20	0.00	5	4	5	2
m-CF_3	0.43	0.20	−0.03	2	2	2	3
m-C_2F_5	0.47	0.20	−0.07	2	2	2	3
p-SCF_3	0.50	0.20	−0.10	1	2	1	1
p-C_2F_5	0.52	0.20	−0.12	5	4	5	2
p-CF_3	0.54	0.20	−0.14	5	4	5	2
m-CN	0.56	0.20	−0.16	3	2	3	4
m-SO_2Me	0.60	0.20	−0.20	4	5	4	5
p-CN	0.66	0.20	−0.26	5	4	5	2
m-NO_2	0.71	0.20	−0.31	4	3	4	5
p-SO_2Me	0.72	0.20	−0.32	7	7	7	7
p-NO_2	0.78	0.20	−0.38	7	6	7	7
m-SO_2CF_3	0.79	0.20	−0.39	4	5	4	5
m-SO_2F	0.80	0.20	−0.40	4	5	4	5
p-SO_2F	0.91	0.20	−0.51	7	7	7	7
p-SO_2CF_3	0.93	0.20	−0.53	7	7	7	7

Table 2. PLS Analysis of Simulated Data Using a Single Property (*P*) or Covariance Matrix (CM), Hodgkin Similarity Matrix (HSM), or Distance Matrix (DM) Calculated from That Property, $n = 31$

(a) Leave-One-Out Cross-Validated Statistics

	PLS model number				optimum no. PLS components				standard error (cv)				R^2(cv)			
data set	*P*	CM	HSM	DM	*P*	CM	HSM	DM	*P*	CM	HSM	DM	*P*	CM	HSM	DM
linear	1	2	3	4	1	1	4	5	0.015	0.001	0.167	0.072	1.00	1.00	0.99	1.00
truncate	5	6	7	8	1	1	2	5	0.415	0.402	0.474	0.047	0.78	0.79	0.72	1.00
optimum	9	10	11	12	1	1	1	5	0.828	0.801	0.826	0.073	−0.16	−0.09	−0.16	0.99

(b) Fitted Statistics

	optimum no. PLS components				standard error (fit)				R^2(fit)			
data set	*P*	CM	HSM	DM	*P*	CM	HSM	DM	*P*	CM	HSM	DM
linear	1	1	4	5	0.014	0.001	0.138	0.037	1.00	1.00	0.98	1.00
truncate	1	1	2	5	0.385	0.385	0.422	0.029	0.81	0.81	0.78	1.00
optimum	1	1	1	5	0.768	0.768	0.764	0.053	0.00	0.00	0.01	1.00

from AM1 partial atomic charges do not completely describe the Hammett σ constants. Figures 3 and 4 show the predictions, again from distance-based PLS, of the artificially constructed nonlinear relationships between electrostatic fields and Hammett σ. Note that the scatter from the theoretical lines is no greater in these plots than in Figure 2 and that there is no systematic lack-of-fit in any region of either Figure 3 or 4.

Table 3. Effect of Adding a Random Variable to the PLS Analysis of the Data Analyzed in PLS Models 4, 8, and 12[a]

	PLS model number		optimum no. PLS components		standard error (cv)		R^2(cv)		standard error (fit)		R^2(fit)	
data set	mixed	separate	mixed	separate	mixed	separate	mixed	separate	mixed	separate	mixed	separate
linear	13	14	8	6	0.076	0.090	1.00	1.00	0.025	0.046	1.00	1.00
truncate	15	16	8	6	0.079	0.061	0.99	1.00	0.026	0.039	1.00	1.00
optimum	17	18	9	5	0.144	0.093	0.98	0.99	0.041	0.070	1.00	1.00

[a] In case 1, mixed, one distance was calculated between each pair of compounds using both the true and random descriptors. In case 2, separate, two distance measures were calculated between each pair of compounds, one for the true and anther for the random variable.

Table 4. PLS Analysis of Simulated Data Based on 49 Benzoic Acid pK_a's Using Electrostatic Fields Calculated in the Region of the Carboxyl at 2 Å Spacing[a]

(a) Leave-One-Out Cross-Validated Statistics

	equation numbers			optimum no. PLS components			standard error (cv)			R^2(cv)		
data set	*F*	CM	DM	*F*	CM	DM	*F*	CM	DM	*F*	CM	DM
linear (Hammett σ)	19	20	21	3	4	8	0.120	0.112	0.107	0.89	0.90	0.92
truncate at $\sigma = 0.20$	22	23	24	3	3	7	0.137	0.139	0.074	0.47	0.44	0.86
optimum at $\sigma = 0.20$	25	26	27	1	1	5	0.198	0.185	0.107	0.14	0.16	0.77

(b) Fitted Statistics

	optimum no. PLS components			standard error (fit)			R^2(fit)		
data set	*F*	CM	DM	*F*	CM	DM	*F*	CM	DM
linear (Hammett σ)	3	4	8	0.105	0.103	0.056	0.92	0.92	0.98
truncate at $\sigma = 0.20$	3	3	7	0.118	0.128	0.039	0.60	0.54	0.96
optimum at $\sigma = 0.20$	1	1	5	0.190	0.183	0.082	0.20	0.18	0.86

[a] The analysis examined the original fields (*F*) as well as covariance matrix (CM) and distance matrix (DM) calculated from these fields.

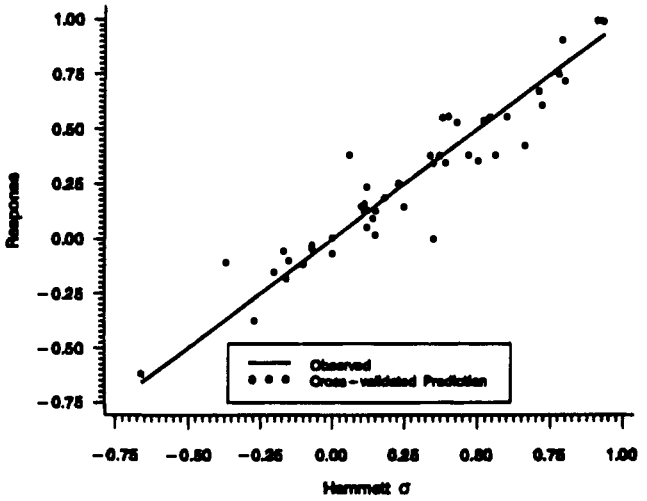

Figure 2. The Hammett σ constant for substituted benzoic acids *vs* the cross-validated fit from distance-based PLS. The Euclidean distances were calculated from the electrostatic field in the region of the neutral carboxyl group, PLS model 21. The cross-validated standard error for this data set is 0.107.

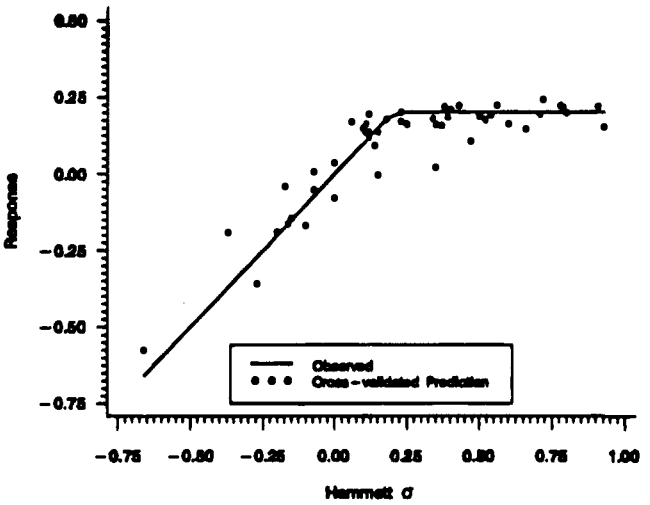

Figure 3. The asymptotic function of the Hammett σ constant for substituted benzoic acids *vs* the cross-validated fit from distance-based PLS, model 24. The Euclidean distances were calculated from the electrostatic field in the region of the neutral carboxyl group. The cross-validated standard error for this data set is 0.074.

Table 5 shows again that the variables used in the construction of the distance matrix must be carefully chosen. For the nonlinear data sets the cross-validated standard errors double if the distance matrices are calculated from both steric and electrostatic fields calculated over the whole molecule, PLS models 31 *vs* 33 and 34 *vs* 36. Note also the decreased statistical quality of the nonlinear models based on distance matrices that include electrostatic fields calculated at points irrelevant to the interaction, PLS models 32 *vs* 33 and 35 *vs* 36. In data not shown, in every case (including Hammett σ) in which there was a model with R^2(cv) > 0.5, the distance matrices were superior to the fields as descriptors.

Lastly, Table 6 includes a comparison of PLS based on Hodgkin and Carbon similarity matrices *vs* PLS based on the original fields or distance matrices. In no case was PLS based on similarities superior or even equal to that based on distance, PLS models 42 and 43 *vs* 44 and 46 and 47 *vs* 48.

From the data in Table 4 we see that the statistics resulting from distance-based PLS provide no guide as to the form of a nonlinear relationship: The cross-validated standard errors from both of the asymptotic and optimum relationships decrease by 46% (PLS models 22 *vs* 24 and 25 *vs* 27) whereas the error from the linear relationship decreases by only 11% (PLS

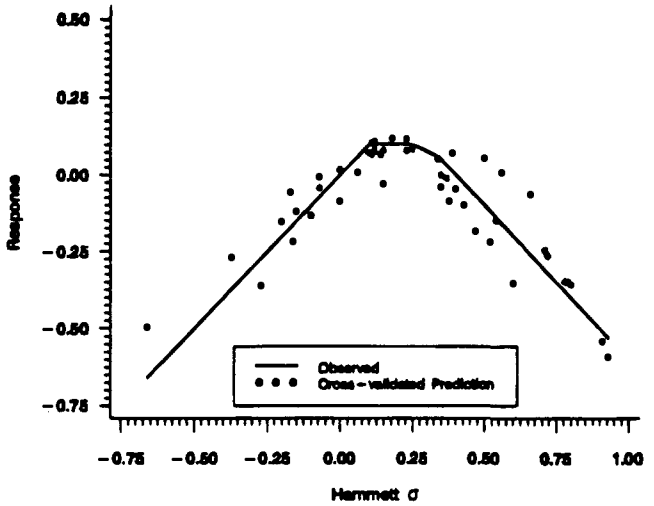

Figure 4. The optimum function of the Hammett σ constant for substituted benzoic acids *vs* the cross-validated fit from distance-based PLS, model 27. The Euclidean distances were calculated from the electrostatic field in the region of the neutral carboxyl group. The cross-validated standard error for this data set is 0.107.

model 19 *vs* 21). Hence, we needed a way to explore the shape of an unknown nonlinear relationship. The simplest is to plot the value of the charge on the atoms involved in the nonlinear relationship identified by PLS.

We clustered the compounds on the basis of their scores in the PLS analyses of the nonlinear relationships listed in Table 4 and also on the basis of their distances to all other compounds. The clustering experiments produced similar results, shown in Table 1. Each cluster represents a set of compounds with similar properties.

If the relationship between a set of descriptors and a target property is linear, then we expect equal variation of the descriptors at the extreme values of the target property: When Table 1 is sorted on the basis of Hammett σ, then the lowest nine (20%) of the compounds are in two clusters and the highest nine compounds are in three clusters. If the relationship is asymptotic, then we expect that there will be more variation in the descriptors at the optimum value of the target property: When Table 1 is sorted on the basis of the asymptotic value, then again the lowest nine are in two clusters whereas the highest nine are in six of the seven possible clusters. Lastly, if there is an optimum value of the descriptor properties, then we expect that there will be more variation in the descriptors at the lower values of the target property: When Table 1 is sorted on the basis of the optimum value, then the lowest nine compounds are in three clusters whereas the highest nine are in only one cluster. Taken together, these observations suggest that a consideration of the number of clusters at each extreme of the dependent property range may also suggest if the relationship is asymptotic or rises to an optimum and falls.

Discussion

Tables 2 and 3 show that the ability of distance-based PLS to detect nonlinear relationships applies to any PLS analysis, not just that involved with CoMFA. Distance-based PLS analysis performs better than both (a) ordinarily regression analysis using a squared term, (b)

Table 5. Analysis of Simulated Data Based on 49 Benzoic Acid pK_a's Using Distance Matrices Calculated from Steric plus Electrostatic Fields Surrounding the Whole Molecule, from Electrostatic Fields Surrounding the Whole Molecule, or from Electrostatic Fields Surrounding the Carboxyl Group Only, 2 Å Spacing

(a) Leave-One-Out Cross-Validated Statistics

	PLS model numbers			optimum no. PLS components			standard error(cv)			R^2(cv)		
data set	both fields, whole molecule	electrostatic, whole molecule	electrostatic, only COOH	both fields, whole molecule	electrostatic, whole molecule	electrostatic, only COOH	both fields, whole molecule	electrostatic, whole molecule	electrostatic only COOH	both fields, whole molecule	electrostatic, whole molecule	electrostatic, only COOH
linear (Hammett σ)	28	29	30	9	10	8	0.145	0.135	0.107	0.86	0.88	0.92
truncate at $\sigma = 0.20$	31	32	33	6	2	7	0.169	0.162	0.074	0.26	0.25	0.86
optimum at $\sigma = 0.20$	34	35	36	3	1	5	0.198	0.156	0.107	0.16	0.40	0.77

(b) Fitted Statistics

	optimum no. PLS components			standard error (fit)			R^2(fit)		
data set	both fields, whole molecule	electrostatic, whole molecule	electrostatic, only COOH	both fields, whole molecule	electrostatic whole molecule	electrostatic, only COOH	both fields, whole molecule	electrostatic, whole molecule	electrostatic, only COOH
linear (Hammett σ)	9	10	8	0.061	0.038	0.056	0.98	0.99	0.98
truncate at $\sigma = 0.20$	6	2	7	0.103	0.146	0.039	0.72	0.39	0.96
optimum at $\sigma = 0.20$	3	1	5	0.164	0.145	0.082	0.43	0.48	0.86

Table 6. PLS Analysis of Simulated Data Based on 49 Benzoic Acid pK_a's Using Electrostatic Fields Calculated over the Whole Molecule (*F*), or the Hodgkin Similarity Matrix (HSM), the Carbo Similarity Matrix (CSM), or the Distance Matrix (DM) Calculated from These Fields, 1 Å Spacing

(a) Leave-One-Out Cross-Validated Statistics

	pls model number				optimum no. PLS components				standard error (cv)				R^2(cv)			
data set	*F*	HSM	CSM	DM	*F*	HSM	CSM	DM	*F*	HSM	CSM	DM	*F*	HSM	CSM	DM
linear	37	38	39	40	6	2	2	8	0.126	0.178	0.177	0.121	0.89	0.75	0.75	0.90
truncate	41	42	43	44	6	2	2	8	0.135	0.159	0.157	0.110	0.52	0.28	0.29	0.70
optimum	45	46	47	48	1	1	1	7	0.187	0.168	0.167	0.160	0.23	0.38	0.38	0.51

(b) Fitted Statistics

	optimum no. PLS components				standard error (fit)				R^2(fit)			
data set	F	HSM	CSM	DM	*F*	HSM	CSM	DM	*F*	HSM	CSM	DM
linear	6	2	2	8	0.051	0.162	0.161	0.029	0.98	0.79	0.80	0.99
truncate	6	2	2	8	0.049	0.144	0.142	0.021	0.94	0.41	0.42	0.99
optimum	1	1	1	7	0.174	0.161	0.160	0.041	0.33	0.43	0.43	0.97

PLS based on the covariance matrix, and (c) PLS based on a similarity matrix. Thus the superiority of the distance-based PLS is not due to simple inclusion of a squared term but is more subtle. We propose that the superiority of the method derives from its ability to model a more complex surface than can be modeled by PLS or ordinary regression analysis. Distance-based PLS is also simpler to run than nonlinear regression since one does not need to supply an equation. The one example provided by Kubinyi showed slightly better cross-validation and fit to distance-based PLS than to nonlinear regression analysis.[8]

Algebraic substitution of eq 1 into eq 3 reveals the relationship between Hodgkin similarity and Euclidean distance:

$$H_{ab} = 1 - \frac{D_{ab}^2}{\sum_k (P_{a,k})^2 + \sum_k (P_{b,k})^2} \qquad (7)$$

Because of the term in the denominator, the similarity between two compounds depends not only on the distance between them, but also where they are located in multidimensional space. At any particular distance value, the similarity can have any value over the allowed range of −1.0 to 1.0; thus the two measures are not correlated.

Why is CoMFA so successful if PLS detects only linear relationships? The reason is that linear relationships are perfectly adequate for shape-related effects since substituents of different shape affect fields at different regions in space. Hence there are energy columns that differentiate large and small substituents. Even nonlinear relationships between potency and hydrophobicity are often accounted for in CoMFA since usually the more hydrophobic analogues are also larger.[17–20] Hence, the detection of nonlinear relationships in CoMFA is critical primarily for electrostatic effects.

A drawback to the use of distance-based PLS analysis is that distance matrices must be carefully constructed if the relationships are to be detected. One might investigate replacing each independent property with a distance matrix or, if the properties are collinear, each principal component of the property matrix could be used to generate a distance matrix. The PLS analysis would be over all distance matrices.

For CoMFA-type fields, we suggest that distance matrices be constructed of electrostatic fields surrounding only a limited region in space, one functional group as shown here. Even if there are five or six such sites in a molecule, there still would be fewer columns in the distance matrices than there were in the original field. The PLS analysis will demonstrate which regions are important. An advantage of calculating distance over small regions of space is that one could decrease the spacing between lattice points while still maintaining interactive performance. In principle, this should improve the quality of the results.

A key advantage of CoMFA over many other 3D QSAR methods is that it leads directly to a graphic display of the results. This advantage is lost with PLS based on distance matrices. Furthermore, the statistics provide no clue if a relationship rises to an asymptote or to an optimum. This is especially difficult when the independent variables are energy fields since there is no simple way to plot potency *vs* field. However, if a result suggests a nonlinear relationship, then as suggested above, the shape of the relationship may be revealed by a simple plot of the value of the charge on the atoms involved in the nonlinear relationship identified by PLS.

There is a subtle problem with traditional cross-validation and distance matrices. Even though one observation is left out by omitting its row, the column corresponding to that compound is typically left in the calculation. Our preliminary analysis suggests that this effect is small. However, it needs to be addressed more thoroughly.

References

(1) Wold, S.; Marten, H.; Wold, H. The Multivariate Calibration Problem in Chemistry Solved by the PLS Method. In *Matrix Pencils (Lecture Notes in Mathematics)*; Kagstrom, R. A., Ed.; Springer-Verlag: Heidelberg, 1983; pp 286–293.

(2) Clark, M.; Cramer, R. D., III. The Probability of Chance Correlation Using Partial Least Squares (PLS). *Quant. Struct.–Act. Relat.* **1993**, *12*, 137–145.

(3) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.

(4) Kubinyi, H., Ed.; *3D QSAR in Drug Design. Theory Methods and Applications*; ESCOM: Leiden, 1993; p 759.

(5) Martin, Y. C. The Quantitative Relationships between pK_a, Ionization and Drug Potency: Utility of Model-Based Equations. In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Ed.; Marcel Dekker: New York, 1980; pp 49–110.

(6) Kim, K. H.; Martin, Y. C. Direct Prediction of Linear Free Energy Substituent Effects from 3D Structures Using Comparative Molecular Field Analysis. 1. Electronic Effects of Substituted Benzoic Acids. *J. Org. Chem.* **1991**, *56*, 2723–2729.
(7) Kim, K. H.; Martin, Y. C. Direct Prediction of Dissociation Constants (pK_a's) of Clonidine-like Imidazolines, 2-Substituted Imidazoles, and 1-Methyl-2-substituted-imidazoles from 3D Structures Using a Comparative Molecular Field Analysis (CoMFA) Approach. *J. Med. Chem.* **1991**, *34*, 2056–2060.
(8) Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*; VCH: Weinheim, 1993; Vol. 1, p 176–177.
(9) Good, A. C.; So, S. S.; Richards, W. G. Structure-Activity Relationships from Molecular Similarity-Matrices. *J. Med. Chem.* **1993**, *36*, 433–438.
(10) Good, A. C.; Peterson, S. J.; Richards, W. G. QSARS From Similarity Matrices - Technique Validation and Application in the Comparison of Different Similarity Evaluation Methods. *J. Med. Chem.* **1993**, *36*, 2929–2937.
(11) Bush, B. L.; Nachbar, R. B. Sample-Distance Partial Least-Squares-PLS Optimized for Many Variables, with Application to CoMFA. *J. Comput.-Aided Mol. Design* **1993**, *7*, 587–619.
(12) Gal, J.-F.; Maria, P.-C.; Chastrette, M.; Zakarya, D.; Exner, O.; Haldna, Ü.; Sjöström, M.; Wold, S.; Zalewski, R. I. Recommendations for Reporting the Results of Principal Component Analysis in the Field of Quantitative Structure Activity Relationships. *Quant. Struct.–Act. Relat.* **1991**, *10*, 52–53.
(13) Efron, B.; Gong, G. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-validation. *Am. Stat.* **1983**, *37*, 36–48.
(14) Stone, M. Cross-validatory Choice and Assessment of Statistical Predictions. *J. R. Stat. Soc., Sect. B* **1974**, *36*, 111–133.
(15) Wold, S. Validation of QSAR's. *Quant. Struct.–Act. Relat.* **1991**, *10*, 191–193.
(16) Goodford, P. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.
(17) Kim, K. H. 3D-Quantitative Structure-Activity Relationships - Describing Hydrophobic Interactions Directly from 3D Structures Using a Comparative Molecular-Field Analysis (CoMFA) Approach. *Quant. Struct.–Act. Relat.* **1993**, *12*, 232–238.
(18) Kim, K. H. Nonlinear Dependence in Comparative Molecular Field Analysis (CoMFA). *Quant. Struct.–Act. Relat.* **1993**, *7*, 71–82.
(19) Kim, K. H. A Novel Method of Describing Hydrophobic Effects Directly from 3D Structures in 3D-Quantitative Structure-Activity Relationships Study. *Med. Chem. Res.* **1991**, *1*, 259–264.
(20) Kim, K. H. Description of Nonlinear Dependence Directly from 3D Structure in 3D-Quantitative Structure-Activity Relationships. *Med. Chem. Res.* **1992**, *2*, 22–27.

JM950196R