A Generalized Formalism of Three-Dimensional Quantitative Structure–Property Relationship Analysis for Flexible Molecules Using Tensor Representation

A. J. Hopfinger, Benjamin J. Burke,[†] and William J. Dunn, III[*]

Laboratory of Molecular Modeling and Design, Department of Medicinal Chemistry and Pharmacognosy, M/C 781, College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois 60612-7231

Received April 29, 1994[®]

A general formalism, based upon tensor representation of multidimensional data blocks, is presented to express relationships between dependent properties and independent molecular feature measures. The solutions to these data set problems are three-dimensional quantitative structure–property relationships, 3D-QSPRs. The molecular features are partitioned into the intrinsic molecular shape tensor, the molecular field tensor, a nonshape/field feature tensor, and an experimental feature tensor. The intrinsic molecular shape tensor contains information on the shape of a molecule within the contact surface while the molecular field tensor contains information outside of the contact surface. Molecular features not directly related to molecular shape are put into the nonshape/field tensor. Experimental measures not being used as dependent variables can be considered as independent molecular features in the experimental feature tensor. The 3D-QSPR is realized by constructing the transformation tensor which optimizes the statistical significance between the dependent and independent variables. Use of partial least squares (PLS) regression permits the unfolding of the composite feature tensor and the identification of the optimum transformation tensor. It is pointed out that a variety of fragment, whole-molecule, two-dimensional, and/or three-dimensional features can be placed into a nonshape/field tensor.

Introduction

A major thrust in computer-assisted molecular design, CAMD, is the construction of three-dimensional quantitative structure–property relationships, 3D-QSPRs. In particular, formalisms are being sought which extract the maximum structure–property information from data sets in which the only initial information components are chemical structures and corresponding property measures.

Most efforts to develop 3D-QSPRs have been directed toward applications in drug design, that is, constructing 3D-QSARs where “A” refers to biological activity and replaces the general assignment of “P” for property. Comparative molecular field analysis, CoMFA,[1] and molecular shape analysis, MSA,[2] are two popular formalisms being used to construct 3D-QSARs. Nevertheless, the concept of relating measures of any common property made for a data set of compounds to calculated physicochemical features of the compounds is the essence of molecular design and not restricted to pharmaceutical applications. QSPRs are, for example, being developed for applications in polymer science.[3] The ability to establish a QSPR for the data set (training set) of compounds offers the opportunity to predict, in advance of synthesis and testing, the value of the property of a compound related, usually an analogue, to those of the training set.

The construction of a 3D-QSPR is a complicated process owing, in part, to the large number of physicochemical features that must be computed and evaluated as correlates to the property measures. However, the large number of degrees of freedom in terms of molecular alignment and conformation, for each possible set of physicochemical features, is mainly responsible for current limitations in 3D-QSPR analysis. Current 3D-QSPR applications generally focus on highly congeneric compounds to limit alignment choices and compounds which are rigid, or are held rigid. In this way the limitations of conformation and molecular alignment can be minimized in constructing the 3D-QSPR. The drawbacks to this approach are to either greatly restrict the universe of compounds that can be tested or to introduce assumptions into the 3D-QSPR model regarding key molecular alignments and conformations for molecular design.

The purpose of this paper is to report the development of a general formalism to construct 3D-QSPRs in which multiple molecular alignments and conformations are considered. Any physicochemical feature can be included in the analysis, and feature selection constraints, to avoid non-real situations, can be imposed. Some attention is also given to the construction of substructure 3D features which represent a new general class of QSPR physicochemical features. An application of the formalism, in terms of analyzing multiple conformer states, is presented for an analogue series of pyridobenzodiazepine inhibitors of muscarinic 2 and 3 receptors in the following paper in this issue.

Methods and Results

1. The MSA–3D-QSPR Physicochemical Feature Tensors. The set of physicochemical features available to explore the construction of a significant molecular shape analysis three-dimensional quantitative structure–property relationship (MSA–3D-QSPR) can be functionally partitioned into three classes:

(1) Intrinsic molecular shape, IMS, features which are usually highly dependent upon conformation. The IMS

[*] To whom inquiries are to be addressed.

[†] Current address: Agouron Pharmaceuticals, Inc., 3565 General Atomics Court, San Diego, CA 92121.

[®] Abstract published in *Advance ACS Abstracts*, September 15, 1994.

0022-2623/94/1837-3768$04.50/0

Table 1. Definitions of Molecular Features and Entities Used To Construct the General MSA−3D-QSPR Formalism

	definitions
u	any compound in the training set, $\{M_u\}$
v	a reference compound
α	the set of conformations for the training set
β	the alignments for the training set
s	the set of intrinsic molecular shape features
p	the set of field probes
$\mathbf{r}_{i,j,k}$	the spatial positions at which the molecular field is evaluated
h_p	the set of non-(molecular shape) and non-(molecular field) features
e_p	the set of experimental measures
f	field-related molecular features not derived from the p

features provide information on molecular shape within the steric contact surface of the molecule.

(2) Molecular field, MF, features which also are highly dependent upon conformation. The MF features provide information on molecular shape beyond the steric contact surface.

(3) The remaining set of physicochemical features which are computed such as lipophilicity, aqueous solubility, conformational entropy, etc., which may, or may not, exhibit a dependence on conformation.

(4) The set of experimental physicochemical features that have been measured for the compounds of interest. These features may, or may not, exhibit conformational dependence. Moreover, any conformational dependence may be realized only as a Boltzmann average for the feature owing to the nature of the experimental measurement. It is also important to note that one or more of these measured properties may, in fact, be used as the dependent variable end-points in the construction of the QSPR.

Table 1 contains a set of definitions to facilitate the formulation of the MSA−3D-QSPR problem for a set of molecules, $\{M_u\}$. Four distinct molecular feature tensors can be constructed from the definitions in Table 1 to incorporate the information associated with each of the four classes of physicochemical features. The IMS tensor is defined as

$$\mathbf{V}_u(s,\alpha,\beta) \text{ or } \mathbf{V}_{u,v}(s,\alpha,\beta) \tag{1}$$

and contains the information regarding molecular shape within the steric contact space. The "u" versus "u,v" representation refers to the use of absolute or relative molecular feature values, respectively. Some molecular similarity measures, s, such as common overlap steric volume (COSV),[4] are measured relative to some reference compound, v. This requires that all molecular features be expressed relative to the corresponding measures for v. Hence, there is the "u,v" form for the IMS tensor. A schematic illustration of the IMS tensor of a single compound is given in Figure 1. It can be seen that the IMS tensor of a single compound is a three-dimensional block of data points in molecular feature, s, conformation, α, and alignment, β, space. The fixed alignment matrix corresponds to a cut through the IMS tensor at particular alignment, $\beta°$, parallel to the (α, s) plane. An example of using the MSA−3D-QSPR formalism for fixed alignment is given in the following paper in this issue. The fixed alignment, fixed conformation vector is also shown in Figure 1 and corresponds to constraints of the CoMFA method.[1] However, in CoMFA the s molecular features are replaced with molecular field features as included in the molecular field (MF) tensor. The IMS tensor for all N compounds in $\{M_u\}$ is schematically illustrated in Figure 2.

The MF tensor is represented as

$$\mathbf{F}_u(p,\mathbf{r}_{i,j,k},f,\alpha,\beta) \text{ or } \mathbf{F}_{u,v} \text{ or } \mathbf{F}_{u,v}(p,\mathbf{r}_{i,j,k},f,\beta,\beta) \tag{2}$$

where $(p,\mathbf{r}_{i,j,k})$ define the set of sampled field potentials and the f are field-related molecular features, such as the dipole moment, not derived from $(p,\mathbf{r}_{i,j,k})$. The composite set $(p,\mathbf{r}_{i,j,k}, f)$ are the "replacements" for s in the IMS tensor and in Figure 1. The MF tensor contains the information regarding the molecular field beyond the steric contact surface of u.

The balance of the molecular features not determined by laboratory experiment are placed in the $\mathbf{H}_u$ feature tensor,

$$\mathbf{H}_u(h_p,\alpha,\beta) \text{ or } \mathbf{H}_{u,v}(h_p,\alpha,\beta) \tag{3}$$

where the h_p are the calculated molecular features not derived from intrinsic molecular shape or molecular field. The h_p may or may not depend on α and/or β, and may be both whole molecule and/or fragment (substituent) based features.

All of the measured physicochemical molecular features used as independent variables are included in the experimental feature tensor,

$$\mathbf{E}_u(e_p,\alpha,\beta) \text{ or } \mathbf{E}_{u,v}(e_p,\alpha,\beta) \tag{4}$$

where the e_p are the experimental molecular feature measures for which an explicit dependence on α and/or β may, or may not, be known.

The dependent variables in the MSA−3D-QSPR training set are organized in the property matrix:

$$\mathbf{P}_u \text{ or } \mathbf{P}_{u,v} \tag{5}$$

where for scalar property measures

$$\mathbf{P}_{u,v} = \mathbf{P}_u \ominus \mathbf{P} \tag{6}$$

Most often $\mathbf{P}_u$ is a vector of the form

$$\mathbf{P}_u = \begin{array}{cc} \mathrm{u} & P_1 \\ \downarrow & P_2 \\ & P_3 \\ & \cdot \\ & \cdot \\ & \cdot \end{array} \tag{7}$$

where P_i is some property measure, such as biological activity of the ith compound in the training set. However, it is also possible that the MSA−3D-QSPR needs to be developed simultaneously for multiple, distinct property measures. For example, in the development of anti-AIDS compounds, via inhibition of reverse transcriptase (RT), it would be best to optimize an inhibitor against the set of observed RT mutations. In such a case $\mathbf{P}_u$ would take the form

$$\mathbf{P}_u = \begin{array}{cllll} & m_1 & m_2 & m_3 & \rightarrow \\ \mathrm{u} & a_{11} & a_{12} & a_{13} & \cdots \\ & a_{21} & a_{22} & & \\ \downarrow & a_{31} & & & \\ & \cdot & & & \\ & \cdot & & & \\ & \cdot & & & \end{array} \tag{8}$$

where the a_{ij} represents the inhibition activities of compound i against the jth mutation. Within the context of developing a general formalism to construct MSA–3D-QSPRs, a general representation of $\mathbf{P}_u$ is

$$\mathbf{P}_u = \begin{array}{llllllll} & P^+{}_{i1} & P^+{}_{i2} & \ldots P^+{}_{i1} & P^-{}_{i1} & P^-{}_{i2} & & \ldots P^-{}_{im} \\ \mathrm{u} & w^+{}_1P^+{}_{11} & w^+{}_2P^+{}_{12}\ldots & w^+{}_nP^+{}_{1n} & w^-{}_1(P^-{}_{11})^{-1} & w^-{}_2(P^-{}_{12})^{-1}\ldots & & w^-{}_m(\mathrm{P}^-{}_{1m})^{-1} \\ \downarrow & w^+{}_1P^+{}_{21} & & & w^-{}_1(P^-{}_{21})^{-1} & & & \\ & \cdot & & & \cdot & & & \\ & \cdot & & & \cdot & & & \\ & \cdot & & & \cdot & & & \end{array} \tag{9}$$

where the superscript (+/−) reflects a positive (+) or negative (−) property, the $w^{+/-}$ are scaling factors to assign different relative weights to the $P_{ik}{}^{+/-}$ and the superscript "−1" indicates that the individual $\mathrm{P}_{ik}{}^-$ may be re-expressed to establish a minimum MSA–3D-QSPR dependence on these property measures. Overall, eq 9 provides a multiple differentiation of property measures.

The most general representation of a MSA–3D-QSPR within the tensor formalism developed here is

Absolute

$$\mathbf{P}_u = \mathbf{T}_u \otimes [\mathbf{V}_u(s,\alpha,\beta),\mathbf{F}_u(p,\mathbf{r}_{i,j,k},f,\alpha,\beta),\mathbf{H}_u(h_p,\alpha,\beta),\mathbf{E}_u(e_p,\alpha,\beta)] \tag{10}$$

Relative

$$\mathbf{P}_{u,v} = \mathbf{T}_{u,v} \otimes [\mathbf{V}_{u,v}(s,\alpha,\beta),\mathbf{F}_{u,v}(p,\mathbf{r}_{i,j,k},f,\alpha,\beta), \mathbf{H}_{u,v}(h_p,\alpha,\beta),\mathbf{E}_{u,v}(e_p,\alpha,\beta)] \tag{11}$$

The quantity in "[ ]" is referred to as the **VFHE** composite tensor. A schematic illustration of eq 10, for which $\mathbf{E}(e_p,\alpha,\beta)$ is "zero" is shown in Figure 3. In eqs 10 and 11 the $\mathbf{T}_u$ and $\mathbf{T}_{u,v}$ are the transformation tensors which optimally map the **VFHE** tensor onto $\mathbf{P}_u$ and $\mathbf{P}_{u,v}$, respectively. The determination of the transformation tensor is the essence of the formalism presented in this paper. The next section discusses current methods being used to establish the transformation tensors.

If a MSA-3D-QSPR, which relates experimental properties and calculated molecular features, is desired, then the property matrix takes the form

$$\mathbf{P}_u = \mathbf{E}_u(e_p,\alpha',\beta') \tag{12}$$

where the prime denotes that the conformation(s) and/or alignment(s) associated with the e_p are not known and/or not needed to measure the e_p. The corresponding form of the MSA–3D-QSPR, in the absolute value representation, is

$$\mathbf{E}_u(e_p,\alpha',\beta') = \mathbf{T}_u \otimes [\mathbf{V}_u(s,\alpha,\beta),\mathbf{F}_u(p,\mathbf{r}_{i,j,k},f,\alpha,\beta),\mathbf{H}_u(h_p,\alpha,\beta)] \tag{13}$$

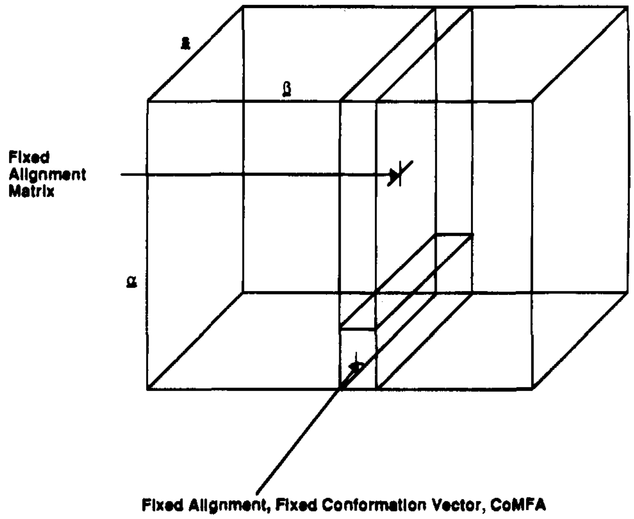

Figure 1. A schematic representation of the IMS tensor for a single compound.

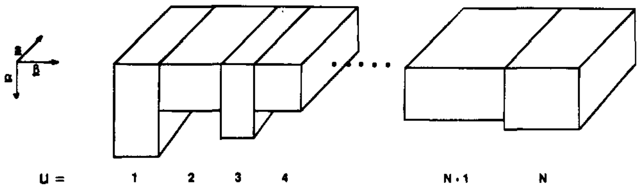

Figure 2. The complete IMS tensor for a set of N compounds presented in schematic form.

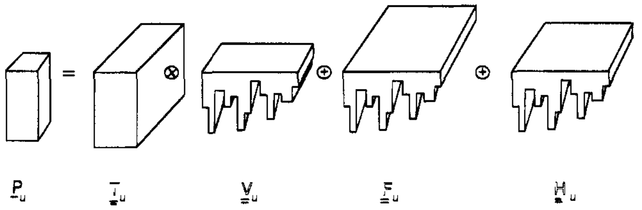

Figure 3. A schematic illustration of eq 11 in which the $\mathbf{E}_u$ tensor is zero.

Equations 10 and 11 incorporate all current QSPR methods as constrained subtypes. For example, CoMFA is given by

$$\mathbf{P}_u = \mathbf{T}_u \otimes [\mathbf{F}_u(p,\mathbf{r}_{i,j,k},f^*,\alpha^*,\beta^*),\mathbf{H}_u(h_p,\alpha^*,\beta^*)] \tag{14}$$

where the asterisk denotes that the degree of freedom is constrained to a single selection. Classic MSA is given by

$$\mathbf{P}_{u,v} = \mathbf{T}_{u,v} \otimes [\mathbf{V}_{u,v}(s,\alpha,\beta^*),\mathbf{H}_{u,v}(h_p,\alpha,\beta^*)] \tag{15}$$

and classic Hansch analysis[5] is given by the reduced form

$$\mathbf{P}_u = \mathbf{T}_u \otimes [\mathrm{H}_u(h_p,\alpha',\beta')] \tag{16}$$

Equations 10 and 11 are completely unconstrained with respect to the selection of the independent variables (degrees of freedom). This introduces the possibility of generating MSA–3D-QSPRs which may not be consistent with known factors. An example of such a situation is the selection of a different molecular alignment in each of the **V**, **F**, and **H** tensors of a single MSA–3D-QSPR. Such a situation would be difficult to realize in an actual system. Hence, there is the need to be able to introduce constraints among the indepen-

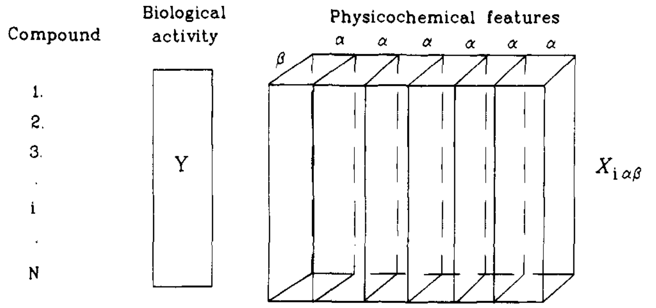

Figure 4. The biological activity matrix and composite 3-way array for a series of compounds. The 3-way array of physicochemical features is constrained so that the same number of conformations and alignments considered are the same.

dent variables. Of course, the fidelity of a 3D-QSPR can be measured, in part, by how plausible it is when no constraints are applied. The introduction of constraints is achieved by using a generalized LaGrangian formalism which, for eq 10, takes the form

$$\mathbf{P}_u = \mathbf{T}_u \otimes \{[\mathbf{V}_u,\mathbf{F}_u,\mathbf{H}_u,\mathbf{H}_u,\mathbf{E}_u] \ominus \lambda\mathbf{C}\} \qquad (17)$$

where λ is the LaGrangian multiplier matrix and **C** is the constraint tensor.

2. Estimation of the Transformation Tensors. The formalism, presented here for the first time, partitions, organizes, and joins all physicochemical features in a common framework—the **VFHE** tensor. However, from the standpoint of treating the degrees of freedom inherent to a MSA–3D-QSPR analysis, it is the consideration of multiple conformations and alignments that make up most of the generalized formalism. Unfortunately, the generalized formalism has corresponding dimensional complications. The large number of degrees of freedom and their multiple representations—the physicochemical feature elements of the **VFHE** tensor—limit the estimation of the transformation tensors in applications where conformational flexibility and multiple alignments are jointly considered. Only components of the general problem embodied by eqs 10 and 11 have been considered to date. The limitations of CoMFA and the current MSA formalism, as expressed by eqs 14 and 15, respectively, relative to eqs 10 and 11, delineate the restrictions inherent to each of these respective methods.

The general statement of the conformation/alignment problem is shown in Figure 4 for N compounds. Here the biological activities are in the 2-way array, **Y**, and the physicochemical features are in the 3-way array, $\boldsymbol{X}$. The 3-way array is a composite of several 3-way arrays, each representing a single feature with each feature a function of conformation, α, and alignment, β. A general solution to the conformation/alignment problem would involve deconvolution of the 3-way array to give the specific conformation and alignment tensors. This would give the **VFHE** conformation and alignment tensors in eqs 10 and 11. We present here only the deconvolution of eq 10 as that of eq 11 is, in fact, a subset, of the solution given.

In the graphical representation of the conformation/alignment problem shown in Figure 4, the measures of each physicochemical feature are elements of a 3-way array in conformation–alignment–compound-space. The net **VFHE** tensor is the composite of the 3-way tables. If the alignment is fixed, the result is a 2-way array delete with each compound expressed as a row vector from each of the 3-way feature tables. The complimentary problem is that in which a single "action" conformation can be identified for all compounds in the data set, but for which multiple alignments are possible.

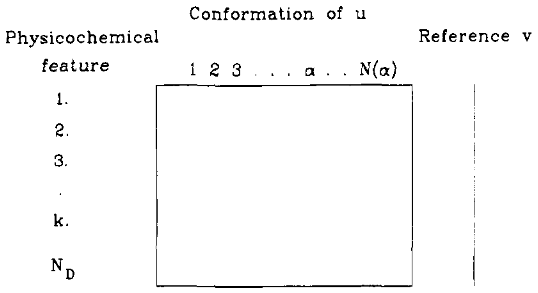

Figure 5. Matrix of physicochemical features for a compound compared to a reference, v.

The application described in the next paper is also limited to a subset of the general solutions inherent to eqs 10 and 11. One important problem not yet considered is that in which the action conformation responsible for expressing the $\{\mathbf{P}_u\}$ can be identified for a reference compound, v. However, other compounds in the series $\{M_u\}$ each can adopt a variety of energetically feasible conformations in which the action conformation is not necessarily of thermodynamic preference. The molecular alignment can be identified, but the problem is to assign the conformations and corresponding physicochemical features to the $\{M_u\}$ which optimize the MSA–3D-QSPR. Treatment of multiple conformations is considered in the following paper.

In order to decide which conformation vector should be used to establish a MSA–3D-QSPR for each u, two assumptions are necessary. First, it must be assumed that the action conformation of a reference compound, v, is known. This requirement can be relaxed in an analysis in that different reference compounds can be tested. The action conformation of the reference compound yielding the optimum MSA–3D-QSPR would be the preferred choice.

The second assumption is that all measures of the physicochemical feature elements of the conformation vectors are equally weighted against one another. This assumption can also be relaxed by assignment of different relative weights to physicochemical features. However, there is usually no basis for making the weighting assignments other than those for optimizing the MSA–3D-QSPR as a function of the weightings.

Overall, given an action conformation for a reference compound, v, and a set of physicochemical feature weightings, the objective now becomes finding the "best" conformer of each u relative to the reference compound, v. Here "best" refers to u being most similar to the reference, v, with respect to the set of physicochemical features. The overall problem for each u can be represented as shown in Figure 5.

Partial least squares (PLS)[6-9] has properties which make it the method of choice to apply in solving the problem represented by Figure 5. PLS will select the u conformation vector which has (1) the greatest variance in terms of the variables, **and** (2) is most highly correlated with the reference conformation vector. It also has the advantage of giving a stable solution when the number of conformations of u approaches or exceeds the number of physicochemical features.

Specifically, PLS is applied to the data in Figure 5 with the reference conformation vector as the dependent, or **Y**, variable and physicochemical feature measures of the conformations of u as the independent, or **X**, data. This maps the **X** data onto **Y** with the constraints discussed above. The PLS model is

$$x_{ik} = \mathbf{x}_i + \sum_{a=1}^{A} t_{ia} z_{ak} + e_{ik} \tag{18}$$

$$y_{ij} = \mathbf{y}_j + \sum_{a=1}^{A} w_{ia} q_{aj} + e_{ij} \tag{19}$$

where $\mathbf{y}_j$ and $\mathbf{x}_i$ are the column means and the t's and w's are latent variables derived from **X** and **Y**, respectively. The z's and q's are the loadings and A is the number of Latent variables computed. The latent variables are related by the inner relation

$$\hat{w} = b^{*}t \tag{20}$$

from which **Y** can be estimated from **X**.

Once the mapping has been done, a criterion for selecting the "best" conformation vector must be established. Cross-validation[10] is used to determine the optimum number of components, A. Cross-validation, which has been discussed more generally,[10] selects the number of PLS components which lead to optimal prediction of the **Y** data.

After establishing the optimum number of components the "best" conformation vector must be obtained. This is done in the following way. The conformation vectors are ranked by deriving an index computed as the sum of the product of the percentage of **Y** variance explained in each component and the square of the **X** loadings. This index is computed over the number of components found to be significant by cross-validation. The index is given by

$$Z_k = \sum_{a=1}^{A} (\% \, \mathbf{Y}_{\text{variance}})(Z_{ak})^2 \tag{21}$$

Identification of the preferred set of conformation vectors permits the generation of a standard QSPR data set. If the number of compounds, N, far exceeds the number of physicochemical features, multidimensional linear regression analysis, as used in traditional MSA, can be employed to establish the MSA–3D-QSPR. On the other hand, if the number of physicochemical features approaches or is much larger than N, as is generally the case in CoMFA applications, PLS is the method of choice to establish the MSA–3D-QSPR.

3. The Physicochemical Features of the H Tensor. There is a tendency to think of the h_p physicochemical features as those from classic Hansch analysis,[5] such as log P and Hammett's σ, connectivity indices like those developed by Kier and Hall,[11] and/or substructural features as developed by Enslein and coworkers for applications in toxicity predictions.[12] These features are largely independent of three-dimensional molecular geometry and can be described as 2D physicochemical features.

Three-dimensional structure calculations provide a means of generating a vast number of 3D physicochemical features which, within the framework of eqs 10 and 11, fall into the h_p class of features. Overall, the h_p physicochemical features for MSA–3D-QSPRs can be divided into four classes.

Class 1: 2D-Substructure Features. Members of this class would include π constants, σ's, and the Sterimol parameters.[13]

Class 2: 2D-Whole Molecule Features. Members of this class would include log P and molecular weight.

Class 3: 3D-Substructure Features. The set of features in this class is limited only by the imagination of the investigator. Often, these physicochemical features are designed by an investigator to probe why some compounds in a data set do not fit a particular 3D-QSPR. The MSA–3D-QSPR study reported in the next paper uses some 3D-substructure features in this fashion. Examples of 3D-substructure features include hydrophobic spheres, dipole moments of individual rings, and conformational entropy of a torsion angle.

Class 4: 3D-Whole Molecule Features. Features in this class can include members of the IMS and MF tensors. However, we formally separate IMS and MF whole molecule features from other 3D whole molecule features. 3D-whole molecule physicochemical features which fall into the h_p class include relative conformational energy, total dipole moment, and solvation energy.

Discussion

The division of the physicochemical features of a molecule into four sets characteristic of intrinsic molecular shape (IMS), molecular field (MF), all other computed physicochemical (h_p), and experimental, measured (e_p) features is arbitrary. The reasons for building this division into the MSA–3D-QSPR formalism are as follows:

1. Members of the intrinsic molecular shape and molecular field feature sets have, respectively, been found to be the major independent variables in many 3D-QSARs.[1,2,14] In fact, it has been proposed[15] that a "starting point" in constructing a 3D-QSAR be a relationship of the form

$$\text{BA} = f(\text{IMS,LIPO,FCDS}) \tag{22}$$

where BA is biological activity, LIPO is some feature related to relative lipophilicity, and FCDS are features characteristic of the data set and usually are needed to describe the behavior of only a few compounds in the data set.

2. Intrinsic molecular shape and molecular field features are, in general, fully dependent upon conformation, alignment, and choice of the measure [e.g., (COSV)/probe (H^+)] used to characterize the feature. Thus, IMS and MF tensors will always require a full exploration with respect to the degrees of freedom.

3. The division permits easy decomposition of the total formalism to explore submodels, such as CoMFA using only the MF and $\mathbf{H}_p$ tensors.

Some $\mathbf{H}_p$ features can be combined with IMS or MF features to construct "higher-order" IMS or MF physicochemical features. For example, the shape commonality index, I_c(u,v),[2] is constructed as a weighted combination of COSV [V_0(u,v)] and the relative intramolecular stability of u, ΔE_u, features in the relative-measure representation of MSA–3D-QSPRs (eq 10)

$$I_c(u,v) = V_o(u,v) - \omega\Delta E(u) \qquad (23)$$

In the construction of such higher-order features one, or more, weighting parameters, like ω, are introduced. These weighting parameters provide a link between diverse physicochemical features. In the case of ω of eq 23, a measure of the shape of a molecule is scaled against its conformational stability. The weighting parameters can be treated the same as the regression coefficients in regression analysis and loadings in PLS, or they can be preassigned prior to the statistical analysis. If the weighting factors are preassigned, then the corresponding data set becomes a hypothesis model relative to the data set with other assigned weighting factors. The statistical analysis comprises the test of the model. Given an arbitrary problem, it is not clear how to formulate an MSA−3D-QSPR analysis in terms of relative or absolute feature measure representation. The relative feature representation guarantees that the compound with the "best" property measure will be a compound to which all others are compared. As such, the corresponding MSA−3D-QSPR should provide a "recipe" for making compounds as good as the "best" one in the data set. On the other hand, the relative feature representation may bias the MSA−3D-QSPR by overweighting the roles of the "better" property compounds in the data set. A particular physicochemical feature which makes a "bad" compound, with respect to the features of the "better" compounds, an "average" compound, may be underweighted, and missed, in the relative feature representation. On the other hand, this feature can appear as a singularity in the "average" compound using the absolute feature representation. Finally, it should be kept in mind that a comparison of the MSA−3D-QSPRs derived from absolute and relative feature representations for a common data set provides information on the stability, significance, and consistency of the QSPR analyses.

Time has not been included as a degree of freedom in the formalism expressed by eqs 10 and 11. Consequently, physicochemical features characteristic of, for example, the time-dependent conformational profile of the molecule are not explicitly included in the MSA−3D-QSPR formalism. However, time-dependent physicochemical features can be incorporated into eqs 10 and 11 by considering time as an additional degree of freedom for each of the tensors. In this case, the data in Figure 3 will be four-dimensional, or 4-way tables.

The repetitive use of PLS is currently the preferred way of identifying optimized solutions to eqs 10 and 11 with respect to conformation, alignment, and concatenation of physicochemical features. PLS permits the maximum correlation between a feature set and a set of property measures. However, the PLS components can contain feature terms that are ambiguous and/or appear to be physically contradictory to other feature terms in the components. For example, two or more conformations, and corresponding physicochemical features, of an analog series of ligands may appear simultaneously as the bioactive conformation in a MSA−3D-QSAR. This clearly leads to questions of optimality and interpretation. Validation procedures can be used to probe the optimality of the model. Also, the PLS model can be expressed in terms of the original variables to generate the equivalent linear regression MSA−3D-QSPRs. A comparison of the regression models to one another, and to the parent PLS MSA−3D-QSPR may provide a direction as to which physicochemical features are most important in the predictive sense. Moreover, the use of genetic algorithms (GA)[16] may be an approach to refining and evaluating the predictiveness of MSA−3D-QSPRs. GA methods have proved very effective in refining some QSARs and QSPRs.[17]

Finally, it is worth remembering that the model derived from a data set can be no better than the component data. The MSA−3D-QSPR physicochemical feature tensor formalism makes no judgement regarding property measures or the set of physicochemical features used. In particular, conformational analysis and molecular alignment require considerable thought and attention as preprocessor steps to generating a MSA−3D-QSPR using the tensor formalism presented here.

In the next paper an application of MSA−3D-QSAR is given using eqs 11, 18, 19, and 21.

Acknowledgment. Resources of the Laboratory of Molecular Modeling and Design at UIC were used as part of this research. B.J.B. gratefully acknowledges fellowships and citation fellowships from the American Foundation for Pharmaceutical Education. Also, the Graduate College of UIC provided fellowship and award support for this work.

References

(1) (a) 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. (b) CoMFA, Tripos Associates Inc., 1699 S. Hanley Road, Suite 303, St. Louis, MO 63144.

(2) Hopfinger, A. J.; Burke, B. J. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.;), Wiley: New York, 1990.

(3) Hopfinger, A. J.; Koehler, M. G.; Pearlstein, R. A.; Tripathy, S. K. Molecular Modelling of Polymers. 4. Estimation of the Glass-transition Temperatures. *J. Polym. Sci. Part B: Polym. Phys.* **1988**, *26*, 2007−2028.

(4) Hopfinger, A. J. A QSAR Investigation of Dihydrofolate Reductase Inhibition by Baker Triazines based upon Molecular Shape Analysis. *J. Am. Chem. Soc.* **1980**, *102*, 7196−7206.

(5) (a) Hansch, C., Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley-Interscience: New York, 1979. (b) Hathaway,G. J.; Hansch, C.; Kim, K. H.; Milstein, S. R.; Schmidt, C. L.; Smith R. N.; Quinn, F. R. Antitumor 1-(X-Aryl)-3,3-dialkyltriazenes. 1. Quantitative Structure Activity Relationships vs. L1210 Leukemia in Mice. *J. Med. Chem.* **1978**, *21*, 563−574. (c) Venger, B. H.; Hansch, C.; Hathaway G. J.; Amrein, Y. V. Ames Test of 1−(X-Phenyl)-3,3-dialkyltriazenes. A Quantitative Structure Activity Study. *J. Med. Chem.* **1979**, *22*, 473−483.

(6) Wold, S.; Wold, H.; Ruhe A.; Dunn, W. J., III, The Colinearity Problem in llinear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM J. Sci. Stat. Comp.* **1984**, *5*, 735−743.

(7) Höskuldsson, A. PLS Regression Methods. *J. Chemomet.* **1988**, *2*, 221−228.

(8) Glen, W. G.; Dunn, W. J., III; Scott, D. R. Principal Components Analysis and Partial Least Squares Regression. *Tetrahedron Comput. Methodol.* **1989**, *2*, 349−376.

(9) Glen, W. G.; Dunn, W. J., III; Sarker, M.; Scott, D. R. UNIPALS: Software for Principal Components Analysis and Partial Least Squares Regression. *Tetrahedron Comput. Methodol.* **1989**, *2*, 377−396.

(10) Wold, S. Cross-validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics* **1978**, *20*, 397−404.

(11) Hall, L. H.; Mahoney, B. K.; Kier, L. B. Electrotopical State Indexes with Molecular Orbital Parameters.-Inhibition of MAO Hydrazides. *Quant. Struct.-Act. Relat.* **1993** *12*, 44−48.

(12) Health Design Inc., Toxicology Newsletter (1993), 183 East Main Street, Rochester, NY 14604.

(13) Franke, R. *Theoretical Drug Destgn Methods*; Akademie-Verlag: Berlin, 1984.

(14) Burke, B. J. Developments in Molecular Shape Analysis to Estimate Spatial Similarity Among Flexible Molecules. Ph.D.

Thesis, Medicinal Chemistry, University of Illinois at Chicago, 1993.

(15) Koehler, M. G.; Rowberg-Schaefer, K.; Hopfinger, A. J. A Molecular Shape Analysis and Quantitative Structure-Activity Relationship Investigation of some Triazine-Antifolate Inhibitors of Leishmeania Dihydrofolate Reductase. *Arch. Biochem. Biophys.* **1988**, *266*, 152−161.

(16) Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley: Reading, MA, 1989.

(17) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation (GFA) to Quantitative Structure-Activity Relationships (QSAR) and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.*, in press.