Perspectives in Drug Discovery and Design, **12/13/14:** 167–182, 1998.
KLUWER/ESCOM

3D QSAR of Flexible Molecules Using Tensor Representation

William J. Dunn III and Antony J. Hopfinger[a]
Department of Medicinal Chemistry and Pharmacognosy, College of Pharmacy, University of Illinois at Chicago, Chicago, IL 60612, U.S.A.

1. Introduction

The process by which a biologically active compound in an *in vitro* or an *in vivo* system is transported and binds to its receptor is poorly understood. This process is an example of molecular recognition [1], and understanding it is a major goal of drug discovery and development research. Computer-aided efforts to understand the process have their beginnings in the early work of Hansch [2], who extended the principles of physical organic chemistry to the study of biological structure–activity relationships. Hansch's work evolved into the field of quantitative structure–activity relationships, or QSAR, which treated drug–receptor interactions as an equilibrium or pseudo-equilibrium process in the same way that substituent effects on the ionization of weak organic acids and bases were treated. The active compounds were quantitatively described by features determined from a consideration of their 2-dimensional structures and these features were correlated with changes in activity. As the appreciation of the role of 3-dimensional structure in biological activity became more acute in the early 1980s, methods of 3-dimensional QSAR, or 3D QSAR, began to emerge. As a note, QSAR studies are a special case of quantitative structure–property relationships, QSPR studies.

In an effort to provide the discussion of 3D QSAR methods with more focus, Hopfinger and Tokarski [3] have recently reviewed this topic and divided the methods into (a) receptor independent and (b) receptor dependent. Receptor-independent methods are developed with little or no prior knowledge of the receptor geometry, while receptor-dependent methods use knowledge of receptor geometry in their derivation. The tensor treatment of structure–activity data to derive 3D QSAR models is a receptor-independent method and is designed to provide information indirectly about the receptor geometry.

By way of introduction to our work, the more important receptor-independent 3D QSAR methods are briefly mentioned here. The reader is referred to the work of Hopfinger and Tokarski [3] for a more in-depth and timely discussion of this topic, and other relevant chapters in this volume.

Tensor analysis has only recently been applied to problems in chemistry. Before its discussion, some definitions and conventions are introduced in order to avoid confusion with terminology. Initially, it is important to distinguish between structural dimensionality and the spatial dimensionality in which the data analysis is carried out. When discussing structural dimensionality, upper-case notation will be used (e.g. 2-Dimensional descriptors or 3D QSAR). Structural Dimensionality is not limited to 3-Dimensions. As

[a] Chem21 Group, Inc., Lake Forest, Illinois, U.S.A.

will be pointed out later in this chapter, the tensor approach encompasses higher structural Dimensions (e.g. time).

The dimensionality of descriptor space will be indicated by lower-case *d* and is determined by the product of the number of descriptors and the number of elements considered in each structural Dimension. For example, if 4 descriptors are evaluated for 10 conformers (conformation is one element of structural Dimensionality) and 15 receptor alignments (alignment is another element of structural Dimensionality), the dimensionality of descriptor space is $4 \times 10 \times 15$.

Tensors are not commonly referred to in computer-aided drug design, even though they are dealt with routinely. For example, a scalar is a zero-order tensor and a vector is a first-order tensor. A first-order tensor is a quantity that has magnitude and direction, while a second-order tensor has magnitude and two directions. Here, column vectors are designated by lower-case, bold characters, **u**. A row or transpose vector is indicated by prime, **u′**. A matrix, or 2-way array, is a second-order tensor and a 3-way array of data is third-order tensor. Matrices are designated as upper-case bold characters, **X**, while 3-way arrays are designated by upper-case, bold italic, ***X***. Higher-order arrays can be represented as *N*-way arrays, where *N* is the order to the tensor. In the social science literature, where tensor analysis is used more extensively, the terminology 2-mode and 3-mode analysis is used. The use of the terminology, *N*-way, is consistent with current usage in the physical science literature and will be used here.

Since a major thrust of the approach presented here is treating structure–activity data of molecules which are conformationally flexible and can assume numerous possible receptor alignments, definitions of conformation and alignment are necessary. Regarding the former, the definition of Eliel et al. [4] is taken: 'By "conformations" are meant the non-identical arrangements of the atoms in a molecule obtainable by rotation about one or more single bonds' [4]. An alignment is the arrangement of two or more molecules in which a common set of atoms, substructures or features is approximately superimposed. In the example presented in this chapter, only pair-wise alignments are used, but the approach presented is not limited to the use of pair-wise alignment rules. The assumption of a reference compound for the pair-wise alignment rule, while a good starting assumption, has limitations. For one, it introduces a bias into the alignment process, and if an error is contained in the reference alignment rule, this error is amplified in the analysis. There would be an advantage, in some cases, in using a 'consensus' alignment rule which is not based on a reference, but gives each compound in the dataset equal weight in the alignment rule. There has been one reference to the use of a consensus alignment rule in structure–activity studies [5], but the method uses an annealing method which is computationally not practical for a large series of compounds.

2. Receptor-Independent 3D QSAR Analysis

Having the 3-dimensional structure of the receptor available to the medicinal chemist reduces drug-design problem to fitting ligands into the receptor site in sterically allowed geometries. While the number of X-ray and nmr determined structures is increasing

rapidly, the majority of drug-design problems require designing ligands for receptors of unknown structure. In such cases, geometric information about the receptor can then be obtained in indirect ways and a number of receptor-independent methods of 3D QSAR have been developed to provide this information.

An underlying assumption of all currently used receptor-independent 3D QSAR methods is that the members of series of bioactive compounds bind to their respective receptor in a common conformation and alignment that allows optimal interaction of the functional groups of the pharmacophore with their complements in the active site.

Comparative molecular field analysis [6,7,8], or CoMFA, is one of the more powerful and frequently used receptor-independent methods. Several other 3D QSAR methods have been proposed and these include molecular shape analysis, or MSA [3], molecular similarity matrices [9], distance geometry techniques [10], the hypothetical active site lattice, HASL, model [11], genetically evolved receptor models, GERM [12], grid analysis [13] and CATALYST [14]. Reference [15] is a good current review of 3D QSAR analysis, and reference [3] provides a focused update and analysis of current work in 3D QSAR. Again, there is no current 3D QSAR approach which is capable of handling the general 3D QSAR problem for flexible molecules for which variable alignment rules can be simultaneously considered. This is the subject of the remainder of this review.

3. The General 3D QSAR Formalism

By relaxing the conformation and alignment constraints imposed by most currently used methods of 3D QSAR, a general formalism for 3D QSAR can be proposed in terms of tensor analysis of the resulting structure–activity data [16]. This formalism is presented here in terms of MSA descriptors. However, in the most general case, it can be applied to any conformation/alignment-dependent descriptor set. The model, in terms of MSA descriptors, is:

$$Y_u = T_u * [V_u(s,m,n), F_u(p,r_{i,j,k},m,n), H_u(h_p,m,n), E_u(e_p,m,n)] \quad (1)$$

where Y is the activity, or dependent variable; conformation is noted by m and alignment by n; and u states that the relationship is absolute rather than relative — i.e. based on a reference compound. In order to use the absolute form of the model, a consensus alignment rule is necessary. The variables, V, F, H and E are four tensors, of which V and F have their roots in MSA. V incorporates shape, s, in molecular description and contains the intrinsic molecular shape, IMS, features of the compounds. It is a measure of the effect of molecular shape within the steric contact surface of the molecule. It is highly dependent on conformation and alignment. F is the molecular field, MF, tensor computed with the set of field probes, p, at spatial positions r_{ijk} from the molecular surface and measures the effect of molecular shape outside the steric contact surface of the molecule. It, too, is highly dependent on conformation and alignment. The H tensor incorporates the physico-chemical descriptors which may or may not be conformation and alignment dependent. Examples are lipophilicity, pK_a, solubility, etc. The E are

largely experimentally determined descriptors for which the conformational dependence is expressed only as a function of the Boltzmann average in the experimental result. The *H* and *E* are the basis of 2-Dimensional QSAR or traditional Hansch analysis and can enter the analysis independently of conformation and alignment. If only information about the geometry of the ligand–receptor complex is of interest, the *H* and *E* may not directly enter into the analysis.

The relative MSA 3D QSAR model is:

$$Y_{u,v} = T_{u,v} * [V_{u,v}(s,a,b), F_{u,v}(p,r_{i,j,k},a,b), H_{u,v}(h_p,a,b), E_{u,v}(e_p,a,b)] \qquad (2)$$

Where the subscript *v* indicates that the tensor is evaluated relative to a reference compound.

The application of the method involves solution for the transformation tensors, T_u and $T_{u,v}$, in Eqs. 1 and 2. The transformation tensors project the descriptors onto the **Y** and can be obtained with a number of data analytical methods. Due to the unique nature of the structure–activity data generally encountered in 3D QSAR, data reduction methods are necessary. Two methods, 3-way factor analysis and 3-way PLS [16] have been applied to this problem and these are discussed below.

3.1. 3D QSAR data structure

The data structure for the 3D QSAR problem with conformation and alignment fixed is shown in Fig. 1. It is identical to the 2-Dimensional QSAR data structure and the data are treated identically. The biological activity measure is **Y**, which is a vector for a single activity or a matrix for more than one measured response. The descriptors, or independent variables, are **X**, and comprise the *V*, *F*, *H* and *E* tensors, as discussed above. In the case of a CoMFA problem, the descriptors are the respective probe-dependent energies computed at points on the grid for each compound. As usual, there are many more variables than compounds, so that a data reduction method — i.e. PLS regression — is required in the data analysis step.

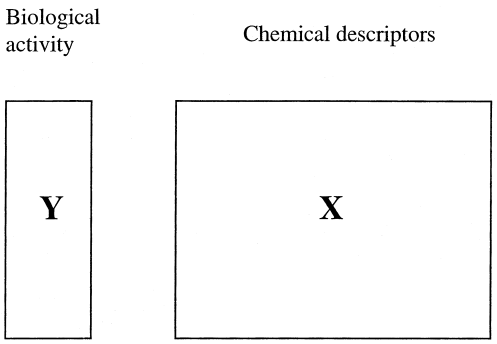


Fig. 1. The standard dataset for a 2D QSAR.

By relaxing the conformation and alignment constraints, the data structure in Fig. 2 results for a single variable. In order to solve the 3D QSAR problem, the resulting 3-way array must be decomposed to yield the transformation tensors, *T*. This can be done in several ways, but the use of 3-way factor analysis and 3-way PLS is proposed. Both have advantages and disadvantages, as will be seen in the discussion which follows.

The use of factor analysis and PLS regression in this application is quite different from their use in traditional 3D QSAR. It is not the objective of their application here to derive a predictive QSAR model, but to solve for the conformation and alignment most highly correlated with activity. It is assumed that only one conformation and alignment is involved in the ligand–receptor complex. However, by varying the resolution of the conformation/alignment space explored and the number of descriptors considered, the 3-way array in Fig. 2 can be small or as large as computationally feasible. It is of interest to extract and rank the important one or two descriptor vectors. These can then be used with more traditional correlation methods, and with other variables, to derive predictive QSARs. In a way, the methods are used here as a variable selector, or filter, to extract the conformation/alignment information from noise.

3.2. 3-way arrays

The QSAR resulting from decomposition of the 2-way array of chemical descriptor data in Fig. 1 provides the change in biological activity with change in 2-Dimensional structure, or with 3-Dimensional structure with conformation and alignment fixed. In the case in which a structure is unconstrained with respect to conformation and alignment, the objective is to decompose the 3-way array in Fig. 2 to explore how the change in structure with respect to changes in conformation *and* alignment is related to the change in biological response. This information is in the unfolded 3-way arrays, as shown in Fig. 3. The unfolding leads to 3 matrices, **O**, **P** and **Q**, which contain the requisite information. The indices *l*, *m* and *n* refer to compound, conformation and alignment,

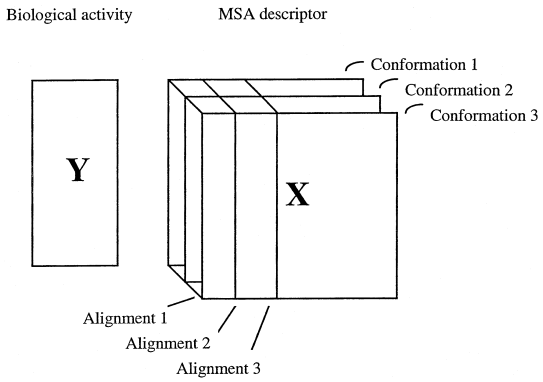


Fig. 2. Data structure for a 3D QSAR problem with one descriptor. Each layer in the x, y-plane of the array represents the variable with conformation fixed and each layer in the y, z-plane represents the layer with alignment fixed. Each column represents the variable with alignment fixed.

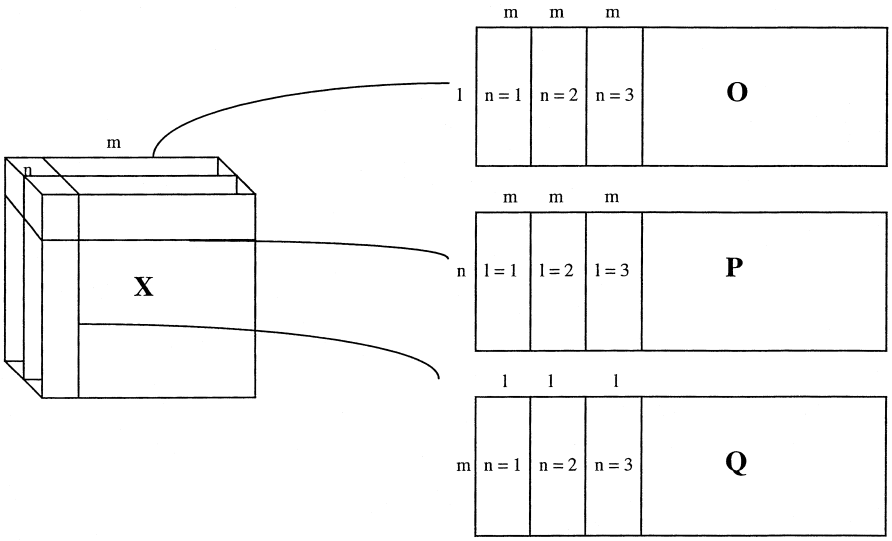


Fig. 3. Unfolding a 3-way array into three 2-way arrays.

respectively, while *o*, *p* and *q* are the number of significant factors or components in the compound, conformation and alignment matrices. 3-Way factor analysis deals with **O**, **P** and **Q**, while 3-way PLS regression deals with **O** from the 3-way array.

3.3. 3-Way factor analysis

3-Way factor analysis was developed first by Tucker [18], and more recently by Kroonenberg [19]. It has also been applied more recently to analysis of analytical [20,21] and environmental chemical [22] data. 3-Way factor analysis decomposes a 3-way array into three factor weight matrices, **A**, **B** and **C**, and a 3-way core matrix, **G** (Fig. 4). The factor weight matrices are associated with compound, conformation and alignment, respectively, with the magnitude of the weights being measures of the variance in the descriptor vectors in the array. The core matrix contains the correlation structure of the 3-way array.

The weight matrices **B** and **C**, which are conformation and alignment specific, are of interest for this application. They indicate the conformation and alignment vectors in the 3-way array which have the greatest systematic variation. The descriptor vectors associated with these heavily loaded conformations and alignments are used in regression to derive the 3D QSAR which is equivalent to principal components regression and subject to the advantages and disadvantages of this method. They are not conditioned to be correlated with **Y**.

The algebraic model for the decomposition is:

$$x_{l,m,n} = \sum_{o}\sum_{p}\sum_{q} a_{l,o} b_{n,p} c_{m,q} g_{o,p,q} \tag{3}$$

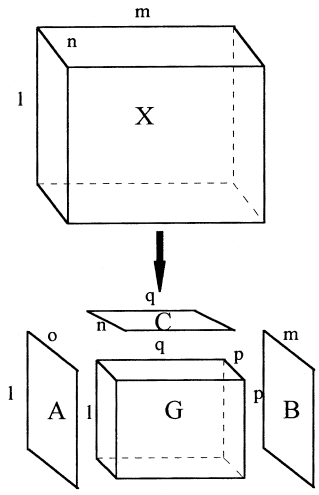


Fig. 4. Decomposition of a 3-way array by 3-way factor analysis.

where *a*, *b* and *c* are the elements of **A**, **B** and **C**, respectively, with *o*, *p* and *q* being the number of significant factors in each. The weights, *o*, *p* and *q*, are not necessarily equivalent. The matrix form is given as:

$$X_{l,m,n} = A_{l,o} G_{o,p,q} (B_{p,n} \otimes C_{q,m}) \tag{4}$$

where the terms are as defined above, and ⊗ indicates the Kronecker product.

3.4. 3-Way PLS regression

Referring to Fig. 5, 3-way PLS regression extracts from ***X*** and **Y** the latent variable which are vectors computed along the axes of greatest variation in ***X*** and **Y** and are most highly correlated. PLS can be applied to ***X*** in terms of a single variable or over a number of variables, *J*. This is shown in algebraic notation in Eqs. 5–7, below. Here, the usual PLS:

$$x_{l,m,n} = \sum_{j=1}^{J} [\bar{x}_{l,m,n} + \sum_{z=1}^{Z} t_{l,z} \otimes P_{z,m,n}] + e_{l,m,n} \tag{5}$$

$$y_{l,i} = \bar{y}_i + \sum_{z=1}^{Z} u_{l,z} d_{z,i} + e_{l,i} \tag{6}$$

$$\hat{u} = b * t \tag{7}$$

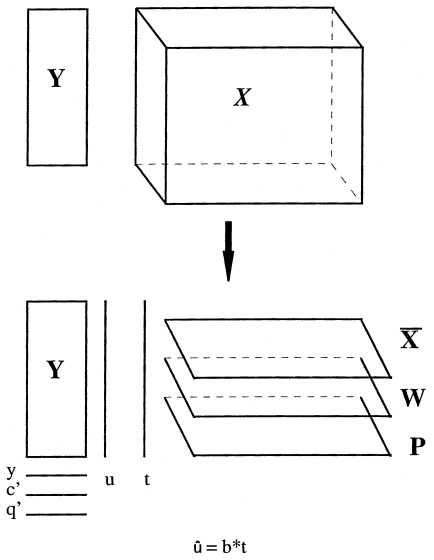
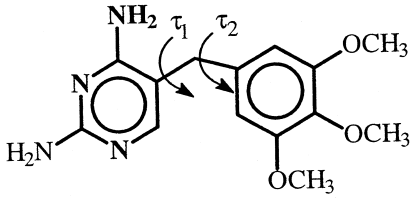


Fig. 5. Decomposition of a 3-way array by PLS regression shown here for one latent variable.

notation is used with *l*, *m* and *n* referring to compound, conformation and alignment, respectively. The latent variables are *t* from the descriptor data and *u* from the biological activity data. The ***X***-loadings are **P** and the **Y**-loadings are *q*. **W** contains the PLS weights. In 3-way PLS, the ***X***-loadings, **P**, are a 2-way array. The number of significant components is *Z*. The sums of the squares of the residuals, $e_{l,m,n}$ and $e_{l,i,}$ are minima. In the calculation of the ***X***-data from the PLS parameters, ⊗ indicates the Kronecker product. Algorithms for computing the 3-way factor and PLS regression models are presented in the algorithm.



I

3.5. Conformation–alignment weights

In order to weight, or rank, the conformations and alignments that result from 3-way PLS, conformation/alignment weights, or CAW, are computed from the ***X***-loadings, **W**; these are computed as below:

$$CAW_{mn} = \sum_{z=1}^{Z} Var_z W_{zmn}^2 \tag{8}$$

Where Var_z is the **Y**-variance explained in component *z*. A similar statistic can be computed from the 3-way factor analysis results by using the sum of squares of the weights from **B** and **C** to rank the conformations and alignments, respectively.

4. Application of the Methodology

In order to illustrate the utility of the 3D QSAR formalism, it has been applied to structure-binding data for trimethoprim, ***I***, and trimethoprim-like analogs to dihydrofolate reductase, DHFR. The geometry of the binary DHFR–trimethoprim complex has been extensively studied [23], making this an ideal set of data for testing the general 3D QSAR formalism. If there is an active conformation and alignment and the tensor analysis approach can predict its geometry, this would help establish its general utility. An account of this work has been published [17], and a summary of the technique and its results are given here.

4.1. Generation of conformation, alignment and MSA descriptor data

Enzyme-inhibitor binding data were taken from the literature on 20 analogs of structure ***I***. Earlier 3D QSAR studies of 2,4-diaminopyrimidine inhibitors of DHFR have shown that the MSA descriptor, common steric overlap volume, COSV, has been a significant variable [24] which led to its use in this study. The structures were built using bond

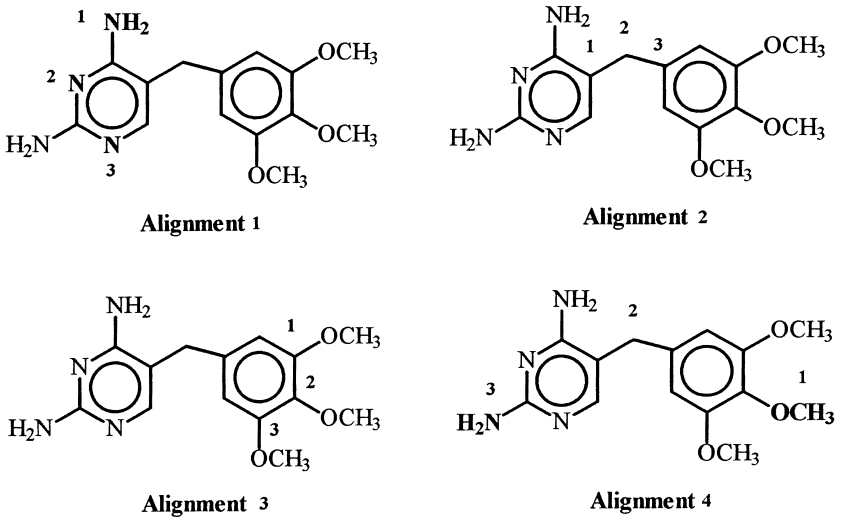


Fig. 6. The four alignment rules used with the key atoms of superimposition noted.

lengths and bond angles from the trimethoprim crystal structure. Partial charges were computed using the MNDO method [25]. Fixed valence conformational analysis was performed for each of the analogs at 10° resolution for the torsion angles, τ_1, and τ_2, as shown in **I**. The MMII non-bonded potential, a Coulomb potential with a dielectric constant of 3.5, and a MMII-scaled hydrogen bonding potential, were used [26]. To be consistent, this force field was used in the study cited above [24]. The conformational profiles of the series of analog inhibitors are defined by the torsion angles τ_1 and τ_2. The conformation of trimethoprim bound in its binary complex with *E. coli* DHFR is defined by torsion angles $\tau_1 = 177°$ and $\tau_2 = 76°$, where $\tau_1 = \tau_2 = 0$ corresponds to the reference conformation in the cis configuration. The active site bound conformation is not the global minimum for any of the analogs. Trimethoprim was used as the shape reference, and 10 trial conformations were considered for each compound. The 10 conformations are operationally equivalent to one another with respect to bonding topology defining the torsion angles, as discussed below.

Trimethoprim is found to have 8 free space minimum energy conformations within 5 kcal/mol of the global intramolecular minimum energy conformation. For each of the other analogs in the dataset, the minimum energy conformations within 5 kcal/mol of the global minimum energy conformation and nearest in τ_1, τ_2 torsion angle space to the minimum energy conformations of trimethoprim were considered; that is, the (10° resolution in τ_1 and τ_2) minimum energy conformations within 5 kcal/mol, closest to the τ_1 and τ_2 values of the selected 8 minima of trimethoprim, were selected. For those compounds that do not have minima for τ_1 and τ_2 values close to those of trimethoprim, the τ_1 and τ_2 values were set to those of the trimethoprim minimum. For the series, overall the τ_1 and τ_2 values vary within a range of ±30° of 177° and 76°, respectively. In total, 10 conformations were selected for each compound, with one conformation being the crystal-bound geometry.

Four alignment rules were selected, as shown in Fig. 6. In each test alignment, 3 key atoms were identified for superposition and all compounds in the dataset are compared pair-wise to trimethoprim using the 3 alignment atoms defining the alignment rule. The COSV for each analog, relative to trimethoprim, for each of the 10 conformations and 4 alignments, was computed. The result was a $20 \times 10 \times 4$ 3-way array. The reader is referred to the original work for further details regarding the structure–activity data. 3-Way factor analysis was applied directly to the 3-way array, and 3-way PLS regression was applied to the data with pIC_{50} as the dependent variable.

4.2. Results

The application of 3-way factor analysis to the data resulted in two significant eigenvalues (based on variance explained) from **M**, **P** and **Q**, respectively. Their eigenvectors were used in the construction of **A**, **B** and **C** (Tables 1–3). The factor loadings were largest for conformation 10, alignment 2, conformation 10, alignment 3 and conformation 9, alignment 2. 3-Way PLS gave results (Table 4) consistent with these with CAW values of 0.10, 0.07 and 0.05, respectively, for the same 3 conformation/alignment sets. The bound conformation of trimethoprim is that of conformer 10, so it is

Table 1 *Factor loadings for the compounds*

Compound	Component number 1	2
1	–0.48	–0.26
2	–0.10	0.01
3	–0.22	0.10
4	–0.07	0.03
5	0.01	–0.08
6	–0.27	–0.08
7	–0.23	–0.07
8	–0.14	–0.11
9	–0.22	–0.16
10	0.13	–0.12
11	0.11	–0.11
12	0.07	–0.06
13	0.47	–0.16
14	0.12	0.07
15	0.30	0.07
16	0.05	–0.05
17	0.26	–0.22
18	–0.09	0.66
19	0.03	0.55
20	0.25	–0.01

Table 2 *Factor loadings for alignments*

Alignment	Factor 1	2
1	–0.53	–0.65
2	–0.18	0.69
3	0.82	–0.24
4	–0.10	0.21

satisfying that the two results give consistent results. Alignment rules 2 and 3 are indicated to be significant in binding and are reasonable in light of nmr spectroscopy studies of the solution structure of the enzyme–inhibitor complex.

To this point, the tensor approach has been used as a filter to extract from the 3-way arrays the geometries of the ligands having the most systematic variation and most highly associated with activity. The descriptor vectors associated with these geometries can be used, either alone or in combination with other descriptors, to develop 3D QSARs. If used with 2-Dimensional structural descriptors, hybrid QSARs result; this is shown below.

The MSA descriptor, $COSV^2$, when regressed with activity gave the 3D QSAR below:

$$\log(1/IC_{50}) = 5.51\,COSV^2 + 2.74$$
$$n = 20,\ r^2 = 0.50,\ XV - R^2 = 0.42,\ F = 17.84,\ F_{1,18\,\alpha\ .01} = 8.28 \qquad (9)$$

Table 3 *Factor loadings for conformations*

Conformation	Factor 1	Factor 2
1	–0.14	–0.40
2	–0.28	0.14
3	–0.24	–0.10
4	–0.32	–0.11
5	0.17	0.52
6	–0.26	–0.09
7	–0.14	–0.31
8	0.21	0.50
9	0.31	0.06
10	0.69	–0.42

Table 4 *PLS loadings for conformation and alignment*

Alignment	Conformation									
	1	2	3	4	5	6	7	8	9	10
1	–0.02	0.00	–0.01	0.00	–0.05	0.00	–0.01	0.00	0.18	0.29
2	0.08	0.03	0.01	–0.03	0.17	–0.02	0.06	0.20	0.01	0.46
3	0.13	0.17	0.14	0.12	0.15	0.13	0.16	0.19	0.23	0.41
4	–0.02	0.04	–0.04	0.02	–0.05	0.01	–0.03	0.03	0.13	0.22

where $XV - R^2$ is the cross-validated R^2 for the equation. The single variable, $COSV^2$, explains 50% of the variation in activity, and when combined with 2- and other 3-Dimensional variables, the result below is obtained:

$$\log(1/IC_{50}) = 0.36\pi - 0.35\,MR^2 + 17.55\,COSV + 0.05\,NOV + 1.89\,S^2 - 10.28 \quad (10)$$
$$n = 20,\ R^2 = 0.91,\ XV - R^2 = 0.82,\ F = 29.33,\ F_{1,14\alpha\ .01} = 3.70$$

where *NOV* is the nonoverlap volume, *S* is the torsion angle unit entropy and *MR* is the scaled molar refractivity.

The tensor analysis approach to 3D QSAR provides computer-aided drug design with a generalized treatment of structure–activity data within a framework of existing QSAR methods. It is an heuristic approach which is subject to the caveats of such methods. The method is based on the same rules of statistics as are all such methods, and in order to be used successfully, they are highly dependent on a good experimental design.

This application indicates the potential for tensor analysis of 3-Dimensional structure–activity to provide information about the receptor-bound geometry of ligands. The methodology is a correlative one and an extension of the 2D QSAR approach. Further applications are under way to explore the utility of tensor analysis not only in 3D

QSAR studies, but in the more general 3D QSAR arena, where it has the potential for providing the structural basis for fundamental processes which have embedded in them complex molecular ordering and orientation.

5. Appendix 5

5.1. Algorithm for decomposition of 3-way arrays by 3-way factor analysis

A variation of the algorithm of Zeng and Hopke [22] has been programmed and is given below:

Step 1. Unfold $\boldsymbol{X}$ to obtain its 3, 2-way arrays, $\boldsymbol{X}_{(l)mn}$, $\boldsymbol{X}_{(n)mn}$ and $\boldsymbol{X}_{(m)ml}$, as in Fig. 3.

Step 2. Compute:

$$\boldsymbol{O}_l = \boldsymbol{X}'_{(l)mn}\,\boldsymbol{X}_{(l)mn}$$
$$\boldsymbol{P}_m = \boldsymbol{X}'_{l(m)n}\,\boldsymbol{X}_{l(m)n}$$
$$\boldsymbol{Q}_l = \boldsymbol{X}'_{lm(n)}\,\boldsymbol{X}_{lm(n)}$$

Step 3. Construct:

$\boldsymbol{A}_{lo}$ from the o significant eigenvectors of $\boldsymbol{O}_l$
$\boldsymbol{B}_{mp}$ from the p significant eigenvectors of $\boldsymbol{P}_m$
$\boldsymbol{C}_{nq}$ from the q significant eigenvectors of $\boldsymbol{Q}_n$.

Step 4. Compute the unfolded core matrix, $\boldsymbol{G}_{(o)pq}$ as:

$$\boldsymbol{G}_{(o)pq} = \boldsymbol{A}_{lo}\boldsymbol{X}_{(l)mn}(\boldsymbol{B}_{mp} \otimes \boldsymbol{C}_{np})$$

Step 5. In the prediction phase, estimate the 3-way array, $\boldsymbol{X}_{lmn}$, where the estimate is in unfolded form:

$$\hat{\boldsymbol{X}}_{(l)mn} = \boldsymbol{A}_{lo}\boldsymbol{G}_{opg}(\boldsymbol{B}_{pn} \otimes \boldsymbol{C}_{qm})$$

Diagnostic statistics can be computed to determine the number of significant eigenvectors, o, p and q, to include in $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$. For this, cross-validation is the method of choice.

5.2. Algorithm for decomposition of 3-way arrays by PLS regression

An algorithm for PLS regression decomposition of 3-way arrays based on the NIPALS algorithm has been published by Lohmöller and Wold [27]. More recently, a cursory discussion of PLS regression decomposition of N-way arrays was published [28], also based on the NIPALS algorithm. Due to the combinatorial problem of treating multiple alignments of flexible molecules, this algorithm is computationally inefficient. Here, a variation of the UNIPALS algorithm [29,30] developed in this laboratory is presented. It differs from the conventional PLS methods, in that it uses a Kronecker product, as does 3-way factor analysis, in the prediction phase. This algorithm has been programmed and, in a limited number of applications, has performed well. Other PLS regression algorithms have been published [31,32] and could possibly be adapted to 3-way array decomposition.

To begin:

Step 1. Compute from $\boldsymbol{X}_{(l)mn}$ and $\boldsymbol{Y}$:
$\boldsymbol{D} = \boldsymbol{X}'_{(l)mn}\,\boldsymbol{Y}$

Step 2. Compute the first eigenvalue, $\boldsymbol{c}$, of $\boldsymbol{D}'\boldsymbol{D}$.

Step 3. Compute the $\boldsymbol{Y}$-scores:
$\boldsymbol{u} = \boldsymbol{Y}\boldsymbol{c}$

Step 4. Compute the $\boldsymbol{X}$-weights, $\boldsymbol{W}$, as:
$\boldsymbol{W}' = \boldsymbol{u}'\boldsymbol{X}_{(l)mn}/\boldsymbol{u}'\boldsymbol{u}$; $\boldsymbol{W}$ is the unfolded form of the 2-way array in Fig. 5. Normalize $\boldsymbol{W}$ to length l.

Step 5. Compute the $\boldsymbol{X}$-scores as:
$\boldsymbol{t} = \boldsymbol{X}_{(l)mn}\,\boldsymbol{W}$

Step 6. Compute the $\boldsymbol{X}$-loadings as:
$\boldsymbol{P} = \boldsymbol{X}'_{(l)mn}\,\boldsymbol{t}/\boldsymbol{t}'\boldsymbol{t}$; $\boldsymbol{P}$ is obtained as the unfolded form of the 2-way array in Fig. 5.

Step 7. Compute the $\boldsymbol{Y}$-loadings as:
$\boldsymbol{q} = \boldsymbol{Y}'\boldsymbol{u}/\boldsymbol{u}'\boldsymbol{u}$

Step 8. Form the inner relation:
$\hat{\boldsymbol{u}} = b^{*}\boldsymbol{t}$

Step 9. Update $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, as:
$\boldsymbol{E} = \boldsymbol{X}_{(l)mn} - \boldsymbol{t} \otimes \boldsymbol{P}'$
$\boldsymbol{F} = \boldsymbol{Y} - b\,\boldsymbol{X}_{(l)mn}\,\boldsymbol{W}\boldsymbol{c}'$

Step 10. To compute the next latent variable, form $\boldsymbol{E}'\boldsymbol{F}$ as the updated $\boldsymbol{X}'\boldsymbol{Y}$ and repeat the algorithm.

In many ways, this algorithm works like regular PLS and the models generated by it can be evaluated in the same way as regular PLS models. In this application, however, the $\boldsymbol{X}$-loadings, $\boldsymbol{P}$, are of interest. The largest elements of $\boldsymbol{P}$ are associated with the receptor-bound conformation and alignment. It may be possible to carry out an orthogonal decomposition of $\boldsymbol{P}$ to obtain the individual conformation and alignment weights but this has not been attempted. Again, cross-validation is the desired method for determining model complexity — i.e. the number of latent variables.

5.3. Kronecker products of matrices

The Kronecker product has not been widely used in the chemical sciences, so that its use may not be familiar to most medicinal chemists. It is used in the prediction phase of both 3-way factor analysis and 3-way PLS. To illustrate its use, consider two matrices $\boldsymbol{Y} = [y_{lm}]$ of order $(i \times j)$ and $\boldsymbol{Z} = [z_{no}]$ of order $(q \times r)$. The Kronecker product, $\boldsymbol{Y} \otimes \boldsymbol{Z}$ will have order $(iq \times jr)$. Unlike the formation of inner and outer products of matrices, the Kronecker product is defined irrespective of the order of the two matrices which are used to form the product. To illustrate the actual operation, consider the two matrices:

$$Y = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \tag{11}$$

$$Z = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \tag{12}$$

The Kronecker product, ($\boldsymbol{Y} \otimes \boldsymbol{Z}$) is:

$$(Y \otimes Z) = \begin{bmatrix} a_{11}Z & a_{12}Z \\ a_{21}B & a_{22}B \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{12}b_{11} & a_{12}b_{12} \\ a_{11}b_{21} & a_{11}b_{22} & a_{12}b_{21} & a_{12}b_{22} \\ a_{21}b_{11} & a_{12}b_{12} & a_{22}b_{11} & a_{22}b_{12} \\ a_{21}b_{21} & a_{21}b_{22} & a_{22}b_{21} & a_{22}b_{22} \end{bmatrix} \tag{13}$$

For further reading the works of Graham [33] and Novotny [34] are recommended.

Acknowledgements

The authors wish to acknowledge the support of the National Science Foundation in the form of a Phase I SBIR grant, and Pfizer Corporation, Groton, CT, U.S.A., in the form of a research grant.

References

1. Roberts, S.M. (Ed.), *Molecular recognition: Chemical and biochemical problems II*, Royal Society of Chemistry, Redwood Press, London, U.K., 1993.
2. Hansch, C., *A quantitative approach to biochemical structure–activity relationships*, Accts. Chem. Res., 2 (1968) 232–239.
3. Hopfinger, A.J. and Tokarski, J.S., *3D-QSAR analysis*, In Charifsom, P.S. (Ed.) Practical applications of computer-aided drug design, Marcel Dekker, New York, 1997.
4. Eliel, E.L., Allinger, N.L., Angyal, S.J. and Morrison, G.A., Conformational analysis, The American Chemical Society, Washington, DC, 1981, p. 1.
5. Barakat, M.T. and Dean, P.M., *Molecular structure matching by simulated annealing: II. An exploration of the evolution of configuration landscape problems*, J. Computer-Aided Mol. Design, 4 (1990) 317–330.
6. Cramer III, R.D., Patterson, R.E. and Bunce, J.D., *Comparative molecular field analysis (CoMFA): 1. The effect of shape on binding of steroids to carrier proteins*, J. Am. Chem. Soc., 110 (1988) 5959–5967.
7. Tripos Associates, 1699 Hanley Road, St. Louis, MO 63144, U.S.A.
8. Cramer, R.D., Clark, R.D., Patterson, D.E. and Ferguson, A.M., *Bioisosterism as a molecular diversity descriptor: Steric fields of single 'topomeric' conformers*, J. Med. Chem., 39 (1996) 3060–3069.
9. Good, A.C., Peterson, S.J. and Richards, W.G., *QSARs from similarity matrices: Technique validation and application in the comparison of different similarity evaluation methods*, J. Med. Chem., 36 (1993) 2929–2937.
10. Crippen, G.M., *Distance geometry approach to rationalizing binding data*, J. Med. Chem., 22 (1979) 988–997.

11. Doweyko, A.M., *The hypothetical active site lattice: An approach to modeling active sites from data on inhibitor Molecules*, J. Med. Chem., 31 (1988) 1396–1406.
12. Walters, D.E. and Hinds, R.M., *Genetically evolved receptor models: A computational approach to construction of receptor models*, J. Med. Chem., 37 (1994) 2527–2536.
13. Goodford, P.J., *A computational procedure for determining energetically favorable binding sites on biologically important macromolecules*, J. Med. Chem., 28 (1985) 849–856.
14. CATALYST, Molecular Simulation, Inc., San Diego, CA, U.S.A.
15. Kubinyi, H. (Ed.), 3D-QSAR in drug design: Theory, methods and applications, ESCOM, Leiden, The Netherlands, 1993.
16. Hopfinger, A.J., Burke, B.J. and Dunn III, W.J., *A generalized formalism for three-dimensional quantitative structure–activity relationship using tensor representation*, J. Med. Chem., 37 (1994) 3768–3774.
17. Dunn III, W.J., Hopfinger, A.J.,Catana, C. and Duraiswami, C., *Solution of the conformation and alignment tensors for the binding of trimethoprim and its analogs to dihydrofolate reductase: 3D-quantitative structure–activity relationships study using molecular shape analysis, 3-way partial least squares regression and 3-way factor analysis*, J. Med. Chem., 39 (1996) 4825–4832.
18. Tucker, L.R., *Determination of parameters of a functional relation by factor analysis*, Psychometrika, 23 (1958) 19–23.
19. Kroonenberg, P., Three mode principal component analysis, DSWO Press, Leiden, The Netherlands, 1983.
20. Apellof, C.J. and Davidson, E.R., *Three dimensional rank annihilation for multicomponent determinations*, Anal. Chim. Acta, 146 (1983) 9–14.
21. Sanchez, E. and Kowalski, B.R., *Generalized rank annihilation factor analysis*, Anal. Chem., 58 (1986) 496–499.
22. Zeng, Y. and Hopke, P.K., *The application of three-mode factor analysis (TMFA) to receptor modeling of scenes particle data*, Atmosph. Environ., 26A (1992) 1701–1711.
23. Koetzle, T.F. and Williams, G.J.B., *The crystal and molecular structure of the antifolate drug trimethoprim (2,4-diamino-5-(3,4,5-trimethoxybenzyl)pyrimidine): A neutron diffraction study*, J. Am. Chem. Soc., 98 (1976) 2074–2081.
24. Mabilia, M., Pearlstein, R.A. and Hopfinger, A.J., *Molecular shape analysis and energetics-based intermolecular modeling of benzylpyrimidine dihydrofolate reductase inhibitors*, Eur. J. Med. Chem.-Chim. Thera., 20 (1985) 163–174.
25. Dewar, M.J.S. and Thiel, W., *Ground states of molecules: 38. The MNDO method, approximations and parameters*, J. Am. Chem. Soc., 99 (1977) 4899–4906.
26. Hopfinger, A.J. and Pearlstein, R.A., *Molecular mechanics force-field parameterization procedures*, J. Comput. Chem., 5 (1985) 486–497.
27. Lohmöller, J.B. and Wold, H., *Three-mode path models with latent variables and partial least squares (PLS) parameter estimation*, In Proceedings of the European Meeting of the Psychometric Society, University of Groningen, The Netherlands, 1980, p. 50.
28. Wold, S., Geladi, P., Esbensen, K. and Öhman, J., *Multi-way principal comonents- and PLS-analysis*, J. Chemometrics, 1 (1987) 41–56.
29. Glen, W.G., Dunn III, W.J. and Scott, D.R., *Principal components analysis and partial least squares regression*, Tetrahedron Comput. Method., 2 (1989) 349–376.
30. Glen, W.G., Sarker, M., Dunn III, W.J. and Scott, D.R., *UNIPALS: Software for principal components analysis and partial least squares regression*, Tetrahedron Comput. Method., 2 (1989) 377–396.
31. Lindgren, F., Geladi, P. and Wold, S., *The kernel algorithm for PLS*, J. Chemometrics, 7 (1993) 45–59.
32. Bush, B.L. and Nachbar Jr., R.B., *Sample-distance partial least squares: PLS optimized for many variables, with application to CoMFA*, J. Comput.-Aided Mol. Design, 7 (1993) 587–619.
33. Graham, A., Kronecker products and matrix calculus: With applications, Ellis Horwood, Chichester, U.K., 1981.
34. Novotny, M.A., *Matrix products with application to classical statistical mechanics*, J. Math. Phys., 20 (1979) 1146–1150.