Molecular Moments for Computer-Aided Drug Discovery

B. D. Silverman*

IBM Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, USA

Abstract: Certain of the fundamental concepts underlying the utilization of comparative molecular moment (CoMMA) descriptors as measures of three-dimensional molecular similarity are reviewed. The results of a principal component regression (PCR) analyses of the five data sets previously examined by partial least squares (PLS) calculations are provided. The results further substantiate the utility of the CoMMA descriptors in predicting chemical and biological activity.

1. INTRODUCTION

The distribution of mass and charge, as well as the shape of a molecule, are properties intimately involved in the delivery and binding of a drug molecule to its targeted receptor site. As a consequence, the three-dimensional representation of these features and their relationship to biological activity has been the subject of numerous investigations [1]. For the case in which receptor site information is not available, one strategy for lead optimization involves a comparison between representations of the three-dimensional molecular features of different molecules. In CoMFA [2] as well as in a number of other procedures detailing three-dimensional molecular fields, molecular comparison requires an alignment or superposition step in which one molecule is oriented with respect to another before molecular comparison can be made. Recently, a number of procedures have been proposed that eliminate the requirement of superposition between molecules [3-6]. The alignment free procedures generate a relatively small set of three-dimensional molecular descriptors, thus enabling greater ease of statistical analysis. This contrasts with the procedures detailing three-dimensional molecular features that result in thousands or more of descriptors that consequently require statistical analyses of under-determined systems such as PLS. Furthermore, for applications requiring the rapid comparison between large numbers of molecules, the alignment free procedures provide an advantage.

The present paper briefly reviews one of these procedures [6], Comparative Molecular Moment Analysis (CoMMA), a procedure that we have developed over the past several years. CoMMA uses the lower order moments of the molecular mass and charge distributions for comparison. A number of such moments, such as molecular weight, inertial moments, and dipole moment, have been previously used in drug discovery procedures. The enhancement of the CoMMA procedure involved the addition of one higher-order multipole moment of the charge density distribution, namely, the quadrupole moment, as well as a description of the relationship between the two distributions by projections of the electrostatic moments upon the principal component inertial axes.

Identification of the center-of-dipole [7], a center defined in an analogous fashion to the center-of-mass or geometric center of the molecule, provides a reference origin for electrostatic multipole comparison beyond the lowest order non-vanishing moment of the charge distribution.

The small number of lower order molecular moments of the mass and charge distributions are a succinct and approximate characterization of these distributions. Even though such moments can be formally defined and calculated approximately, it wasn't clear initially that they would contain a sufficient amount of information to correlate with either chemical or biological activity. Subsequent investigation [8], however, discovered significant correlation between the calculated moments and the activities of several molecular series previously investigated by other procedures.

In most cases the small number of CoMMA descriptors enables one to perform QSAR's for over-determined systems, i.e., for systems in which the number of training set molecules is greater than the number of descriptors plus offset of the regression equation. This provides the opportunity of utilizing principal components simply in examining how well the descriptors correlate with biological and chemical activity and consequently to obtain well-defined statistical measures of how good the results of the regression might be.

The present review consists of a discussion of several of the concepts underlying the CoMMA procedure as well as a reexamination, by principal component regression (PCR), of the five chemical series that had been examined previously [6] by PLS.

The next section, section 2, describes the development and derivation of the electrostatic multipole moment descriptors. Section 3 provides results of the PCR performed on the molecular series and the final section consists of a brief summary.

*Address correspondence to this author at the IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, USA; Tel: (914)-945-1135; Fax: (914)-945-4104; Email: silverma@us.ibm.com

1389-5575/01 $20.00+.00

2. MOLECULAR MOMENTS

The molecular charge consists of contributions from two components; the distribution of electron charge and the positively charged atomic nuclei. Even though the net charge might be equal to zero, the distribution of charge is responsible for an electrostatic field that is present exterior to the molecule. Such field can be represented in a number of different ways. One way is by an infinite series, the leading terms of which contribute a greater-and-greater percentage to the total field as the position that one examines becomes farther-and-farther removed from the molecule. Each term of the series involves an "electrostatic moment" which is a particular characterization of the molecular charge. The first term of this series contributes to the field as would a point charge with magnitude and sign just equal to the net molecular charge. The second term involves the familiar dipolar contribution the field.

Terms of increasing order of the series involve contributions to the electrostatic field in increasing angular variation. With $\rho(\vec{r})$, the molecular charge density at position $\vec{r}$ measured from the origin of the coordinate system, the first three lowest order electrostatic molecular moments are written:

zeroth order: net molecular charge

$$q = \int \rho(\vec{r})\, d^3r$$

For a molecule with a non-vanishing net charge, this term makes a contribution to the electrostatic field that is spherically symmetric.

The first order moment is the dipole:

$$\vec{p} = \int \vec{r}\rho(\vec{r})\, d^3r$$

The angular variation of the dipolar contribution to the field exhibits sign reversal upon a rotation of 180°.

The second order moment is the quadrupole:

$$\overline{\Theta} = \frac{1}{2}\int (3\vec{r}\,\vec{r} - |\vec{r}|^2\,\overline{1})\rho(\vec{r})\, d^3r$$

Certain terms of this second order contribution to the field can be represented by a four-leaf clover with sign alternating upon a 90 rotation.

So each increase in multipole moment order, characterizes the increase in angular variation that the charge distribution makes to the electrostatic field. For locations that are not distant from the molecule, the series might be slowly convergent and the first few terms provide a poor approximation to the total electrostatic field. The first few terms still, however, provide the complete characterizationof a particular angular variation of the field about the origin of expansion that has been chosen.

For molecules with net charge equal to zero, the zeroth order multipole moment vanishes and the lowest order non-vanishing multipole moment, if non-vanishing itself, is the dipole. All molecules presently considered will be assumed to have a non-vanishing dipole moment. This excludes molecules such as benzene from present consideration. For such set of molecules with net charge equal to zero and non-vanishing dipole moment, only the dipole moment will have a value that does not depend upon the origin of the coordinate system about which it is calculated. All higher-order molecular moments, e.g., quadrupoles, octapoles, hexadecapoles,…, will have values that depend upon the location of the origin of the expansion. Consequently, whenever one sees a dipole moment descriptor assigned to a molecule in any drug discovery procedure, such assignment has been made for a molecule with zero net charge, namely, a neutral molecule.

One can define a special center, the center-of-dipole, as the origin of the multipolar expansion in a manner somewhat analogous to the definition of the center-of-mass. Recall that the center-of-mass is the location of the origin of expansion of the mass distribution for which the first order moment vanishes. Since the first order moment of the charge distribution is the lowest order non-vanishing multipole moment, by analogy one might attempt to find a location for the electrostatic multipolar expansion for which the next order moment or quadrupolar contribution vanishes. While this is possible for a linear molecule [9], it is not possible to zero out all five independent quadrupolar components by a displacement in three-dimensional space. One can, however, find the origin of expansion for which the quadrupolar components make a minimal contribution to the averaged far electrostatic field. This origin is formally obtained by minimizing the solid angle average of the squared deviation between the dipolar field and total field [7].

For a multipolar expansion about this center, one then rotates to the principal quadrupolar set of axes, namely, the set of axes for which the quadrupolar tensor (three by three matrix) is diagonal. In this Cartesian reference frame one finds a particular relationship between the direction of the dipole and one of the principal quadrupolar coordinate axes; namely, they lie along the same direction, (Fig. **1**). Furthermore, the diagonal element of the matrix associated

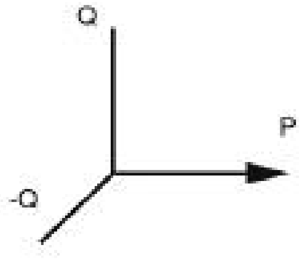

Fig. (1). Principal Quadrupolar Cartesian frame of reference with dipole moment, P, and principal quadrupole moment, Q.

with this direction is zero. Since the matrix is traceless, the sum of the diagonal elements adding up to zero, the two non-vanishing diagonal elements must be equal in magnitude and opposite in sign. This provides one additional electrostatic multipole moment descriptor for neutrally charged molecules, namely, q, the magnitude of the two non-vanishing quadrupolar principal values. Such descriptor, together with the dipole, depends upon molecular structure only implicitly, through the determination by such structure of the electrostatic distribution of charge. So an electrostatic descriptor has been defined by analogy with the moments-of-inertia which depend upon the definition of the center-of-mass. The quadrupole descriptor depends upon the definition of the center-of-dipole.

Together with these two electrostatic descriptors, p and q, one also has moments of the distribution of the molecular mass, namely the molecular weight, m, and moments of inertia, I_x, I_y, I_z, for a total of six descriptors. Eight additional descriptors can be introduced that relate the charge to the distribution of mass, namely, the magnitudes of projections of the dipole upon the principal inertial axes, p_x, p_y, p_z, and two components of the quadrupolar tensor written in the frame of the principal inertial axes, $\boldsymbol{q}_{XX}$ and $\boldsymbol{q}_{YY}$. These five descriptors, together with magnitudes of the displacements between the center-of-dipole and center-of-mass complete the following fourteen descriptors that have been previously utilized.

$$p_X \; p_Y \; p_Z \; p \; q \; q_{XX} \; q_{YY} \; I_X \; I_Y \; I_Z \; d_X \; d_Y \; d_Z \; m$$

The values of these descriptors can be calculated in a number of different ways [8]. Calculation of the moments of the mass distribution is straightforward. Since most of the molecular mass is concentrated at the nuclei, it is sufficient to just perform sums involving mass weighted nuclear positions. Calculation of the moments of the charge distribution may well be straightforward but accompanied by uncertainty as a consequence of near cancellations between contributions from the positively charged nuclear cores and the negatively charged electronic charge distribution. Electrostatic molecular moments can be obtained from ab-initio quantum chemistry calculations about an arbitrary origin. With values of the dipole moment and quadrupole moments one can then translate to a coordinate frame with origin positioned at the center-of-dipole. After calculation of the quadrupolar moments about the center-of-dipole, the quadrupole moment descriptor is then obtained by the rotation to principal axis orientation.

The electrostatic moments can also be calculated from point charges assigned to the atomic centers. For molecules of any reasonable size, all such calculations are expected to yield moments with considerable error, however, if the calculations yield systematic variations that mirror actual trends, this can provide important correlative information for the QSAR regressions performed.

The fourteen descriptors are correlated and for descriptor elimination it is convenient to deal with an uncorrelated or orthonormal set of principal component descriptors. The relatively small number of descriptors enables one to systematically eliminate components that correlate marginally or not at all with the observed or measured activity. Principal components can be eliminated from the regression that results in a marginally decreasing r^2 and consequent increase in F value. With such procedure only those descriptors that correlate best with the activity data are retained. Such strategy of descriptor retention differs from the strategy that retains only those principal components responsible for the major variance of the data. This latter strategy might eliminate important components of reduced variance that correlate significantly with the data. The reduced variance of such components might well be a consequence of the arbitrary normalization procedure adopted for descriptors of mixed chemical and physical character and it is therefore best to retain such components in the regression.

3. PRINCIPAL COMPONENT REGRESSION (PCR) OF THE FIVE MOLECULAR SERIES

The five chemical series have been examined previously [6] by a partial least squares (PLS) analysis. The present discussion reexamines these series by principal components regression (PCR). The PCR procedure also provides information concerning the statistical significance of the results.

The five series reexamined are: 1. Twenty-one steroids with corticosteroid binding affinity data. 2. Fifteen substituted imidazoles with pKa data. 3. Forty-nine substituted benzoic acids with Hammett constant data. 4. Thirty-seven β-carboline and related molecules with binding affinity data for the benzodiazepine receptor site. 5. Thirty-three anti-HIV TIBO derivatives. Details of the molecular structural determination and calculation of the moments have been given previously [6].

(Table **1**) shows results obtained by performing the PCR analysis on the series. Principal components have been obtained by diagonalizing the correlation matrix of the original set of correlated descriptors. Values of r^2 and the Fisher statistic, F, are given when all principal components have been used in the analysis. These are the same values one would obtain by performing the regression with the original set of correlated descriptors. The number of components with regression coefficients having a 95% confidence interval that does not intersect zero is also provided. So, for example, for the calculation with all fourteen steroid components, there were four principal component regression coefficients, well defined, with 95% confidence intervals removed from zero. All calculations have been performed with the MATLAB Statistics Toolbox [10]. Under, “optimized F”, the strategy adopted has been to eliminate components that do not contribute to an increase in the F statistic. Listed under “comp”, is the number of components required to achieve a maximal F value. No attempt was made to optimize r^2. Also, of the ten components retained for the steroids, five had regression coefficients with 95% confidence intervals that did not intersect zero.

Calculations have also been performed for what has been called, “slipped data”, in which the activity data has been rearranged with the uppermost entry placed at the bottom of

Table 1.

	Steroids	**Imidazoles**	**Benzoics**	**Carbolines**	**Tibo's**
all components					
r^2	0.991	0.989	0.839	0.676	0.763
F	44.1	15.24	12.6	4.17	4.14
95%	4	0	5	3	2
optimized F					
r^2	0.988	0.985	0.815	0.576	0.519
F	80.5	37.1	25.7	14.9	10.4
comp	10	9	7	3	3
95%	5	3	5	3	1
Slipped Data					
all components					
r^2	0.619	0.885	0.555	0.498	0.291
F	0.695	1.28	3.03	1.98	0.526
95%	0	0	1	0	0
optimized F					
r^2	0.37	0.442	0.213	0.131	0.131
F	3.33	10.3	12.7	5.27	2.26
comp	3	1	1	1	2
95%	0	0	1	0	0

the list. The second entry now appears at the top of the activity list of data and all other entries are also translated up by a single position. With such translation of the activity data there is no longer the correct registration between the molecular activity data and the descriptors of the molecule for which the measurement has been made. Scanning the results of (Table **1**). One sees that the results achieved with the correct registration between the activity and molecular descriptors for all of the series are significantly degraded when the calculations are performed with the “slipped data”.

4. SUMMARY

The present paper has briefly reviewed several of the concepts underlying comparative molecular moment analysis (CoMMA). It has also exhibited the significant correlation between the moment descriptors and the chemical and biological activities of several chemical series previously investigated by other procedures. Such correlation has also been observed for several other molecular series not presently described. Since the three-dimensional moment descriptors encapsulate certain fundamental aspects of molecular shape, mass, and charge, as well as their relationship, perhaps this is as should have been expected. The small number of these descriptors coupled with their lack of an explicit superposition requirement provides a convenient and succinct three-dimensional representation for the assignment of molecular similarity.

REFERENCES

[1] See, for example, 3D QSAR in Drug Design Volumes 1-3, edited by H. Kubinyi *et al.* (1993/98) Kluwer/Escom, Dordrecht, The Netherlands.

[2] Cramer, R. D. III; Patterson, D. E.; and Bunce, J. D. *J. Am. Chem. Soc.*, **1988**, *110*, 5959.

[3] Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T.W. *Journal of Computer-Aided Molecular Design*, **1999**, *13*, 271.

[4] Bursi, R.; Dao,T.; van Wijk, T.; de Gooyer, M.; Kellenbach, E.; Verwer, P. J. *Chem. Inf. Comput. Sci.*, **1999**, *39*, 861

[5] Todeshini, R.; Gramatica, P. in 3D QSAR in Drug Design, H. Kubinyi, G. Folkers, Y. C. Martin Ed.; Kluwer/Escom Dordrecht, The Netherlands, 1998; Vol. 2, pp. 355-380.

[6] Silverman, B. D.; Platt D. E. *J. Med. Chem.*, **1996**, *39*, 2129.

[7] Platt, D. E.; Silverman, B. D.; *J. Comput. Chem.*, **1996**, *17*, 358.

[8] The web site, http://www.research. ibm.com/comma includes a list of publications of previous work utilizing the CoMMA descriptors.

[9] Buckingham, A. D. in Advances in Chemical Physics, J. O. Hirschfelder Ed, John Wiley & Sons, **1967**, Vol. *12*, 107-143.

[10] MATLAB Statistics Toolbox, *The MATHWORKS Inc.* 24 Prime Park Way, Natick, MA 10760.