

Validation of DAPPER for 3D QSAR: Conformational Search and Chirality Metric

Scott A. Wildman[†] and Gordon M. Crippen*

College of Pharmacy, University of Michigan, 428 Church Street, Ann Arbor, Michigan 48109

Received September 19, 2002

Adequate conformational searching of small molecules and inclusion of a chirality identifier are necessary features of any current technique for quantitative structure–activity relationships (QSAR). However, implementation of these features can be difficult and computationally expensive, and some techniques can still lead to insufficient treatment of molecular conformation. We select the standard systematic conformational search as the default search method for our recent 3D QSAR program, DAPPER, and develop a novel chirality metric for use in QSAR. These techniques are implemented in DAPPER and validated on standard data sets.

INTRODUCTION

Many of the molecular modeling methods used in the pharmaceutical industry and elsewhere rely on a common set of fundamental computational techniques. For instance, conformation searching^{1,2} is common to many methods for pharmacophore modeling, ligand docking, and three-dimensional quantitative structure–activity relationships (3D QSAR). Diversity measures³ are used in library design, high-throughput screening analysis. Chirality metrics⁴ are found in molecular similarity calculations, QSAR, and enumeration of isomers. Improvements to two of these techniques will be discussed here. As the underlying techniques develop, the field for each of the methods can also develop, and often the leading methods or programs are those that can easily incorporate new underlying techniques and improve results without significantly changing the method.

Our recent 3D QSAR method, DAPPER, was developed with this in mind and therefore can easily incorporate different conformation search techniques and chirality metrics. Here we discuss the analysis of several implementations of various conformation search techniques, a novel chirality metric, and their incorporation into DAPPER with validation on three standard medicinal chemistry data sets. The programs described in this work are available from the corresponding author.

METHODS

DAPPER Summary. The DAPPER method is described in detail in a recent publication,⁵ and only a brief summary will be provided here. DAPPER generates a QSAR model from a typically small set of molecules with known biological activity using descriptors from atomic physicochemical property data and intramolecular atom pair distances. The natural progression of a recent method from Crippen,⁶ DAPPER, is based on the idea that tight-binding ligands seek to place similar property values in similar positions in the environment of the binding protein⁷ and avoids many of the

pitfalls of earlier methods while maintaining several recent improvements. DAPPER models are of deliberately low resolution, yet accurately predict biological activity over a range of structures and activity values.

Three-dimensional structure information is incorporated using intramolecular atom pair distances where the distances used are not based on any single conformation of a molecule but rather are the minimum and maximum distance possible for each atom pair. The set of distance ranges for all atom pairs in a molecule is determined by conformation search, a topic discussed in detail in this work. Further, the set of distance ranges is divided into segments in order to eliminate infeasible distance combinations resulting from correlated atom movement. The result of this complex treatment is that each molecule is not represented as a single conformation, or even a set of discrete conformations, but rather the entire feasible conformation space is represented.

Physicochemical property information is included using the SLOGP and SMR atomic contributions to partition coefficient and molar refractivity⁸ and Gasteiger–Marsili partial charges.⁹ These property and distance range data are categorized using a variable resolution, histogram-like set of Gaussian functions for each property, and only representative descriptor values are used to construct the QSAR model using a modified partial least squares algorithm. The number of Gaussian functions used to represent each property is variable, and as a result, the number of descriptors also changes. Additionally, descriptors for molecular chirality may be included using a method described below.

The activity data values, usually K_i or IC_{50} , are treated as intervals to best represent the imprecise nature of many such measurements. Equivalent to error bars, for each molecule, m , this interval is $g_l < g_{obs} < g_u$, where g_l and g_u are the lower and upper limits on the error bar interval, respectively. Due to this treatment of experimental data, the commonly used r^2 and q^2 metrics become meaningless. Instead, for a molecule to fit the model or be accurately predicted, we require $g_l < g_{calc} < g_u$.

The data set is divided into two disjoint subsets: a training set used to generate QSAR models and a test set used to assess the predictive ability of each model. Typically, the

* Corresponding author e-mail: gcrippen@umich.edu.

[†] Present address: Pfizer Global Research and Development, Ann Arbor, MI.

training set is comprised of only a few active molecules to start with, but DAPPER moves molecules between the training and test sets as it runs, although at any time the subsets are disjoint. The size of each training set is not constant across biological systems, nor is it a set fraction of the overall data set.

As the program runs, a model of deliberately low specified resolution, a low "level" model, is generated to fit the data in the training set using a variation⁶ of PLS. The "level" is presented as [D,H,R,Q] where D, H, R, and Q are the number of Gaussians used to represent distance, logP, refractivity, and charge, respectively. If no model can be fit at this low resolution, the level of resolution is gradually increased until a model can be found to fit the data. When a model is found that fits data in the training set, the test set is then used to verify the predictive ability of the model. If a given model can both fit the data in the training set and accurately predict the biological activities in the test set, the model is accepted. In this manner, the lowest resolution model is found. Since DAPPER fits a model to intervals of activity rather than specific values, it is possible to determine more than one equivalent low resolution model, but each model must meet the same fitting and prediction criteria and on that basis each model is either in full agreement with the data or is rejected altogether.

Chirality. The importance of chirality in molecular recognition is undeniable, and this section presents a simple continuous chirality measure developed for use in 3D QSAR methods. The general notion of chirality has intrigued many people in mathematics, physics, and chemistry, resulting in a huge literature of sometimes great mathematical sophistication that is neither feasible nor appropriate to review here. In chemical applications alone, the concept has been refined in several different ways in order to address certain experimental issues, such as optical activity and isomerism, or to extend to general theories of molecular similarity, molecular symmetry, or enumeration of isomers. No one viewpoint, however mathematical and elaborate, completely addresses the needs of all possible applications.

This technique is yet another novel treatment of molecular chirality that is better suited to identifying whether the molecule is achiral, and if not, quantitatively how chiral. A similar technique was developed independently by Moreau,¹⁰ but as a major difference, that method assigns chirality values to individual atoms, rather than to the molecule as whole.

Given atomic coordinates for a particular conformation of a molecule and some property value assigned to each atom, it provides an easy way to calculate a chirality value, χ , that distinguishes enantiomers, is zero for an achiral molecule, and is a continuous function of the coordinates and properties. In DAPPER, this value will be determined using each of the three atomic physicochemical properties (SLOGP, SMR, and charge) and atomic mass. The goal is a chirality measure that maps a particular conformation of a particular molecule to a real number, and this mapping must have the following properties. (1) While it may depend on conformation, it must be independent of overall rigid translation and proper rotation of the molecule. (2) Mirror reflection or improper rotation of the molecule in that conformation must change the sign of the result but not the magnitude, implying that achiral molecules are mapped to zero. (3) The chirality measure must depend on some atomic

(or group) property appropriate to the application at hand, rather than on distinguishability rules that are fixed for all time. (4) The measure should be a continuous function of conformation and property values.

This measure is based on a standard result of classical mechanics: if a configuration of particles has a plane of symmetry, i.e., a reflection plane, then this plane is perpendicular to a principal axis where a principal axis is defined to be an eigenvector of the inertial tensor. Furthermore, if the configuration of particles possesses any axis of symmetry, then this axis is also a principal axis, and the plane perpendicular to this axis is a principal plane corresponding to a degenerate principal moment of inertia.^{11,12} In other words, there are two tests for an achiral molecule: it either has a reflection symmetry in a plane perpendicular to a principal axis, and/or its inertial tensor has a degenerate eigenvalue.

The standard conversion from an arbitrary coordinate system to principal axes goes as follows. Each atom i has given Cartesian coordinates \mathbf{a}_i and mass m_i . Then translation to center-of-mass coordinates is just

$$\mathbf{c}_i = (c_{xi}, c_{yi}, c_{zi})^T = \mathbf{a}_i - (\sum_j m_j \mathbf{a}_j) / (\sum_j m_j)$$

where the superscript T indicates matrix transpose. From this we calculate the inertial tensor \mathbf{I} .

$$\mathbf{I} = \begin{bmatrix} \sum m_i (c_{yi}^2 + c_{zi}^2) & -\sum m_i c_{xi} c_{yi} & -\sum m_i c_{xi} c_{zi} \\ -\sum m_i c_{xi} c_{yi} & \sum m_i (c_{xi}^2 + c_{zi}^2) & -\sum m_i c_{yi} c_{zi} \\ -\sum m_i c_{xi} c_{zi} & -\sum m_i c_{yi} c_{zi} & \sum m_i (c_{xi}^2 + c_{yi}^2) \end{bmatrix}$$

Any such real, symmetric matrix has eigenvectors that can be chosen to be normalized, mutually orthogonal, ordered by increasing eigenvalue, and the last one multiplied by -1 if necessary to form a right-handed coordinate system. Let \mathbf{R} be the 3×3 matrix whose rows are these ordered eigenvectors. Then the principal axis coordinates of each atom are just $(x_i, y_i, z_i)^T = \mathbf{R} \mathbf{c}_i$.

The new chirality measure is defined in terms of the principal axes coordinates by

$$\chi = \begin{cases} 0 & \text{for degenerate eigenvalues} \\ \sum m_i x_i y_i z_i & \text{nondegenerate} \end{cases}$$

which satisfies all four desired properties. The conversion to principal axes coordinates ensures independence of translation and proper rotation, because the conversion is uniquely determined when \mathbf{I} has nondegenerate eigenvalues. The eigenvectors are not uniquely determined when two or more eigenvalues are equal, but in that case, achirality has already been detected, and $\chi = 0$. For the second property, observe that a reflection operation, σ , corresponds to changing the signs of all x_i , say, which reverses the sign of χ , but not the magnitude. A nonplanar but achiral molecule will be positioned so that one principal axes plane, for instance the yz -plane, is the plane of symmetry, and each atom i either lies on that plane ($x_i = 0$) or there is a matching atom j across the plane such that $m_i = m_j$, $y_i = y_j$, $z_i = z_j$, and $x_i = -x_j$, thus producing $\chi = 0$. Although in classical physics the inertial tensor is calculated by weighting the particles by their masses, one can identify m_i with any atomic property of

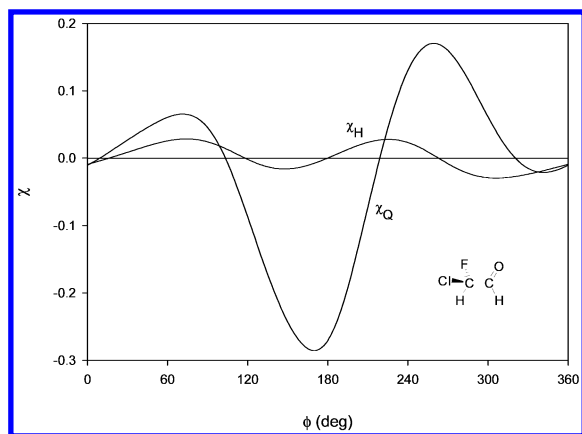


Figure 1. Conformational variability of χ_H and χ_Q .

interest, thus satisfying property 3. Finally, the fourth property is satisfied by the case of nondegenerate eigenvalues and by the continuity of the principal axes as a function of the atomic masses and original coordinates. A perturbation from the degenerate case to the slightly nondegenerate case positions the molecule such that at least one of the principal axes is nearly an axis of symmetry, resulting in a near zero value of χ .

For the purposes of comparison between molecules of different sizes, it is useful to use a normalized metric as

$$\chi_{\text{scaled}} = \frac{10^5 (\lambda_2 - \lambda_1)(\lambda_3 - \lambda_1)(\lambda_3 - \lambda_2)}{\lambda_1 \lambda_2 \lambda_3 \sum_i m_i} \sum_i \frac{m_i x_i y_i z_i}{(x_i^2 + y_i^2 + z_i^2)^{3/2}}$$

where $\lambda_1 \leq \lambda_2 \leq \lambda_3$ are the eigenvalues of the inertial tensor. This version is independent of the size of the molecule and of the magnitude of the weights, in that multiplying the coordinates or the weights by a constant positive factor leaves χ_{scaled} unchanged.

This chirality description is unique to each conformation of a given molecule, and while it is not generally useful to list the χ values of a long list of molecules, it is appropriate to illustrate a few examples. A mostly conformationally rigid example of a chiral molecule is *R*-fluorochloromethanol. For a given conformation, while the H-C-O-H bond rotation is *trans*-, $\chi_M = -0.0002$, $\chi_H = 0.3865$, $\chi_R = 0.0311$, and $\chi_Q = 0.0981$ for the chirality values determined using the properties of atomic mass, SLOGP, SMR, and atomic partial charge, respectively. For the same conformation, $\chi_{M\text{scaled}} = -104.0$, $\chi_{H\text{scaled}} = 104.4$, $\chi_{R\text{scaled}} = -271.9$, and $\chi_{Q\text{scaled}} = 138.8$. Inverting the chiral carbon to create the *S*-isomer results in opposite sign, same magnitude results for all properties ($\chi_M = 0.0002$, $\chi_H = -0.3865$, $\chi_R = -0.0311$, and $\chi_Q = -0.0981$).

With a slightly larger example, *R*-fluorochloroacetaldehyde, χ clearly has a geometric component, and it comes as no surprise that it varies as a function of conformation as can be seen in Figure 1. Shown are χ_H and χ_Q as the Cl-C-C-H dihedral angle is varied through its full range. Here the magnitudes of χ are slightly larger. As can be seen, there are several conformations at which each of the properties produces $\chi = 0$; however, these do not occur simultaneously. This feature holds as long as the properties chosen for χ

evaluation are not well-correlated. Having $\chi = 0$ for a certain property is analogous to the experimental fact that the optical rotation of a chiral compound can be zero at certain wavelengths.

DAPPER was developed in the Molecular Operating Environment (MOE)¹³ which is equipped to determine chirality using the Cahn-Ingold-Prelog (CIP)¹⁴ chirality rules. MOE assigns a value of 1 or -1 for chiral molecules and zero for achiral molecules. The main drawback of this approach is the difficulty involved with correct identification of CIP priorities. The Moreau atomic chirality relies solely on atom properties and position, thus avoiding this difficulty, but assigns a chirality value to each atom. While this might seem to fit well with the DAPPER atomic property based descriptors, it should be noted that the Moreau method does not necessarily assign high chirality values to the chiral centers of a given molecule. In fact, it seems to be the norm that achiral atoms will have quite high chirality values.

Our new chirality measure is derived from only atom positions and clearly quantifiable properties, in this case SLOGP, SMR, charge and mass, avoiding the CIP priority rules. It provides a quantitative description of the overall chirality of a given conformation of a molecule without assigning values to individual atoms. Quite suitable for use in QSAR and molecular modeling applications, it is simple to calculate and can accommodate any atomic property values. For these reasons it has been included in DAPPER as the default chirality measure.

Conformation Search. Incorporation of conformational flexibility in QSAR methods is vital,¹⁵ and the distance range segments used to accomplish this in DAPPER can be arrived at using any number of conformational searching methods. However, the segments can only accurately represent the conformational space of the molecule if the search algorithm pushes all atom pair distances to their extremes, which requires a thorough conformation search. Such a search should consistently identify the same distance range minima and maxima over multiple runs on the same molecule(s).

Several search algorithms were investigated for reproducibility using 10 "drug-like" small molecules and a set of six steroids, each with 2-8 rotatable, nonterminal bonds, 0-4 flexible rings, and molecular weight <500. These molecules were chosen from standard QSAR data sets and were expected to be a reasonable representation of the kinds of molecules encountered in a drug discovery process.

The algorithms investigated included a random search in Cartesian coordinates¹⁶ (atom position), a random search in internal coordinates¹⁷ (torsion angles), and a systematic search of torsion angles, all standard searches in MOE. Also tested was a biased random search in Cartesian coordinates in which a penalty function is evaluated during optimization in order to generate unique conformations, in this case, those with unique sets of atom pair distances, that was based on the idea of poling.¹⁸

A conformation search for each molecule was carried out 10 times with each algorithm, and the resulting conformations were compared to assess the reproducibility of the method. Only the systematic search successfully generated the same list of conformations for repeated searches with the same molecule, while the other algorithms often found less than half of the same conformations in separate runs even with liberal stopping criteria and long search times.

Since the steroids used in the investigation have published crystal structures, it was also possible to determine which algorithms could reproduce these experimental conformations. The systematic search correctly reproduced the crystal structure coordinates within a small tolerance while the other search methods consistently failed to match the experimental geometry for more than four of the six steroids, even when given multiple attempts.

Other search methods available in MOE were tested to a lesser degree, with similar irreproducible results. Only the systematic search was reproducible and could consistently find crystal structure coordinates. The systematic search is, however, significantly more computationally expensive than the random search methods, prohibitively so for molecules with many flexible rings or rotatable bonds. Nonetheless, the systematic search must be used, since any conformational search method that gives irreproducible results cannot be used, no matter what speed benefits it may provide. Such a search will be known to not always identify the global energy minimum conformation. In the current application, an inconsistent search algorithm will not necessarily provide consistent range segments and can therefore result in different QSAR results with the same molecules.

In the current version of DAPPER the MOE systematic search forms the root of the range segment search procedure. However, the modular nature of DAPPER will easily allow new conformational search algorithms to be used as they become available.

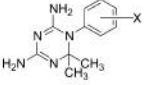
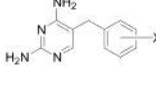
RESULTS AND DISCUSSION

DAPPER results need to present several values in order to be complete. Each solution can be comprised of several different models at a given level of resolution, and therefore each set of results needs to define the level and provide the QSAR coefficients for each model. The level is presented as [D,H,R,Q], where D, H, R, and Q are the number of Gaussian functions used to represent the distance, SLOGP, SMR, and charge property values, respectively. Other values of interest are the number of molecules used in the training and test sets, the number of descriptors used, and number of PLS vectors needed for an acceptable model. Manual inspection of the coefficient values is not terribly descriptive, yet these values are the main result of DAPPER and are used to predict biological activity data for previously unknown molecules.

While long run times for the generation of these models are typical, it is important to consider the complexity of the task. DAPPER does not simply produce a list of descriptors and solve the system using PLS as many QSAR programs do. Instead it is searching for potentially multiple models to fit interval data using a complete representation of conformation space at variable resolution and meeting strict acceptance criteria. The utility of DAPPER was shown in its original presentation,⁵ and here three standard QSAR data sets will be used to illustrate the incorporation of updated conformational search and chirality methods.

Dihydrofolate Reductase. The standard dihydrofolate reductase (DHFR) data set is comprised of 23 4,6-diamino-1,2-dihydro-2,2-dimethyl-1-(substituted phenyl)-5-triazines,¹⁹ 24 2,4-diamino-5-(substituted phenyl)pyrimidines,²⁰ and methotrexate shown in Table 1, compounds **1**–**48**. The

Table 1. DHFR Activity Data^a

Triazines (1 – 23 , P1 – P9)				Pyrimidines (24 – 47 , P10 – P20)			
							
No.	X—	log <i>K</i> _{calc}	<i>g</i> _{calc}	No.	X—	log <i>K</i> _{calc}	<i>g</i> _{calc}
		(log <i>K</i>)				(log <i>K</i>)	
1	H	−4.70	−5.17	24	H	−5.20	−5.71
2	3-I	−5.18	−4.71	25	3-OBu	−6.13	−6.13
3	3-OBz-3',4'-Cl ₂	−5.57	−5.77	26	3-I	−6.67	−6.00
4	4-OCH ₃	−4.10	−4.51	27	3,4,5-(OCH ₃) ₃	−6.88	−6.65
5	3-SO ₂ NH ₂	−2.93	−3.12	28	3-F	−5.38	−5.75
6	3-COCH ₃	−4.24	−4.39	29	3-CH ₂ OH	−5.67	−5.72
7	3-OH	−3.87	−3.47	30	4-NH ₂	−5.47	−5.18
8	3-CF ₃	−4.77	−5.15	31	3,5-(CH ₂ OH) ₂	−5.73	−5.70
9	3-F	−4.88	−4.52	32	4-F	−5.67	−6.06
10	3-CN	−5.31	−5.52	33	3,4-(OH) ₂	−5.84	−6.22
11	3-CH ₃	−4.96	−4.47	34	3-OH	−5.82	−5.25
12	3-CH ₂ CH ₃	−5.40	−4.71	35	4-CH ₃	−5.83	−5.41
13	3-OCH ₃	−4.52	−4.83	36	3-CH ₂ OBu	−5.49	−5.50
14	3-OCH ₂ CH ₃	−5.19	−5.71	37	3-CH ₃	−5.78	−5.85
15	3-OPr	−5.58	−5.27	38	4-OCH ₃	−6.25	−6.74
16	3-OHx	−5.69	−5.62	39	4-OBu	−6.37	−6.44
17	3-OBz	−5.68	−6.20	40	4-NHCOCH ₃	−6.05	−4.28
18	3-CH ₂ OPh	−6.57	−5.91	41	3-OCH ₃	−5.93	−6.47
19	4-OH	−4.91	−5.26	42	3-OBz	−6.19	−6.19
20	4-NH ₂	−3.94	−3.63	43	3-CF ₃	−6.16	−5.84
21	4-I	−4.43	−4.74	44	3-CF ₃ , 4-OCH ₃	−7.30	−6.66
22	4-CH ₃	−4.17	−4.59	45	3,4-(OCH ₃) ₂	−6.92	−6.56
23	4-F	−4.65	−4.18	46	3,5-(OCH ₃) ₂	−6.42	−5.82
				47	3,5-(OH) ₂	−3.38	−3.71
				48	methotrexate	−9	−13.4
P1	4-COCH ₃	−3.52	−3.59	P10	3-CH ₂ OCH ₃	−5.64	−6.17
P2	4-CF ₃	−3.68	−4.30	P11	3-Cl	−5.90	−5.53
P3	4-Cl	−4.76	−4.71	P12	3-OCH ₂ CONH ₂	−5.96	−6.72
P4	4-OBz	−5.19	−5.84	P13	4-OCH ₂ CH ₂ OCH ₃	−6.05	−6.93
P5	4-CH ₂ SPh-3'-CH ₃	−5.55	−6.51	P14	3-OCH ₂ CH ₂ OCH ₃	−6.12	−6.63
P6	4-CH ₂ SPh	−5.61	−6.45	P15	4-N(CH ₃) ₂	−6.17	−7.34
P7	4-SBz	−5.64	−6.12	P16	4-OCF ₃	−6.30	−5.74
P8	3-SBz	−6.00	−5.85	P17	4-OBz	−6.35	−6.08
P9	3-CH ₂ OPh-3'-t-Bu	−6.45	−6.49	P18	4-Ph	−6.41	−6.05
				P19	3-OCH ₃ , 4-OH	−6.47	−6.27
				P20	3-OH, 4-OCH ₃	−6.59	−6.50

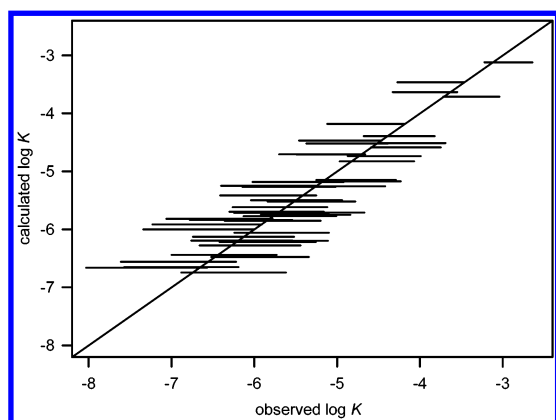
^a Binding data used as interval log *K* ± 10%.

experimental activity intervals for these conformationally flexible molecules were taken as [*g*_i, *g*_u] = log *K* ± 10%. The structures from this data set have been reproduced extensively in the literature, and will not be in this work. While a variety of data is presented in the original works, the biological activity data used here, and presented in Table 1, is taken entirely from inhibition of *L. casei* DHFR.

The range segment search was completed using the systematic search, and overall, the process took just over 25

Table 2. DHFR DAPPER Model

level	set training molecules	test set molecules	PLS adjusted DAPPER coefficients
[1,1,2,4]	1 2 3 4 5 6 7	9 14 25 28	-24.1482 0-136.473-153.272 222.374 0
	8 10 11 12	29 32 37 38	-175.126 73.9301 51.4252-570.457 383.725
	13 15 16 17	39 43	3819.61-89.3071 0-1209.31 562.998 664.569
	18 19 20 21		0 0 108.014-1152.18 0 0 0-341.94 0-147.212
	22 23 24 26		420.543-363.626 0-1343.05-89.7328 0 0 0 0
	27 30 31 33		-2940.67 0 0 0-1291.31 0-521.06-21683.7
	34 35 36 40		642.564 0 0 1816.67 0 0 44.3476 2585.49
	41 42 44 45		7951.52 0 0 0 2677.1 0 0 963.119-172.403 0 0
	46 47 48		0 0

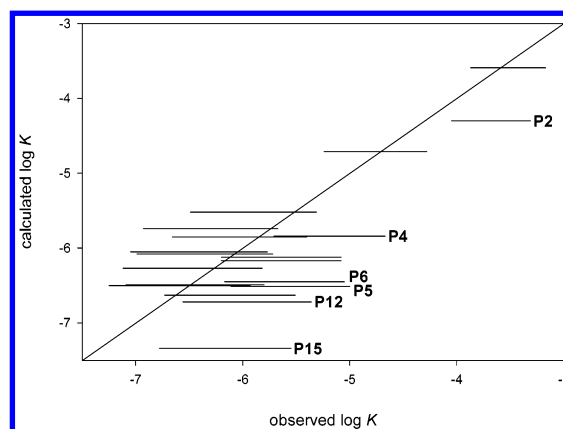
**Figure 2.** Correlation plot for DHFR Inhibitors. The data are represented as line segments over the experimental interval.

h on a SUN Ultra 60 for the 48 molecule data set, for an average of near 30 min to perform the range segment search on each molecule.

A single DAPPER model was found at level = [1,1,2,4] based on 31 PLS vectors from 65 descriptors using 38 molecules in the training set and 10 in the test set, as presented in Table 2, in 39 h. The calculated values of all molecules are presented in Table 1. Additionally, a plot of observed vs calculated activity is presented in Figure 2. Since observed activities are intervals, the data are horizontal line segments, and those that cross the diagonal line representing observed = calculated are correct, in-range predictions.

Comparison of these results with other methods is difficult due in part to the uniqueness of this approach. For example, the original studies by Hansch and co-workers^{19,20} considered the pyrimidines separately from the triazines and analyzed several subsets of each group. For their two-dimensional QSAR models they report $r^2 = 0.413$ to 0.790 and $\sigma = 0.340$ to 0.214 for the pyrimidines and $r^2 = 0.291$ to 0.923 and $\sigma = 0.710$ to 0.244 for the triazines. While these statistics are not used in DAPPER, the final model found has an $r^2 = 0.968$ and $\sigma = 0.411$ when comparing the calculated values to the midpoints of the experimental activity intervals. Additionally, Hansch notes that in the original work there are differences in those features identified as being important when their QSAR model is run on binding data from different species and even on data from the same species in a whole-cell assay rather than a purified enzyme assay. This may suggest that there are multiple modes of action for the given sets of inhibitors. Whether or not this is the case, it would be inappropriate to consider $r^2 = 0.291$ a reasonably fit model.

A true prediction of DHFR inhibitors was performed on a separate set of nine triazines¹⁹ and 11 pyrimidines^{20,21} presented in Table 1 as compounds **P1**–**P20**. Following the

**Figure 3.** Correlation plot for prediction of DHFR Inhibitors. Compound numbers are presented for those not predicted in-range.**Table 3.** DAPPER Calculated ACE Inhibition Values^a

no.	log IC ₅₀	g_{calc} (log IC ₅₀)	no.	log IC ₅₀	g_{calc} (log IC ₅₀)
1	-6.1	-6.140-6.254-6.143	16	-7.0	-7.483-7.401-7.298
2	-7.4	-8.152-8.163-8.154	17	-8.6	-9.073-8.785-7.902
3	-6.0	-5.399-5.401-5.432	18	-7.4	-8.138-8.141-8.140
4	-8.4	-9.255-9.154-9.273	19	-7.3	-6.628-6.783-7.052
5	-8.2	-8.612-7.477-7.505	20	-7.7	-7.923-7.572-7.279
6	-8.0	-8.102-7.541-8.084	21	-8.9	-9.185-8.425-8.033
7	-8.8	-7.972-8.033-8.079	22	-8.5	-8.815-9.311-8.595
8	-8.5	-9.256-8.478-9.123	23	-9.0	-9.688-9.784-8.783
9	-9.0	-9.662-8.252-9.893	24	-9.6	-8.675-8.679-8.676
10	-8.1	-8.626-7.422-8.727	25	-8.5	-8.971-8.035-7.818
11	-7.6	-7.239-7.358-6.901	26	-8.4	-9.077-9.238-8.791
12	-8.9	-9.272-9.452-8.119	27	-7.9	-7.632-7.160-7.255
13	-8.0	-7.244-7.246-7.245	28	-8.8	-9.037-7.994-8.888
14	-9.2	-9.626-10.14-9.836	29	-4.0	-3.039-3.295-1.744
15	-8.5	-9.404-9.386-9.188	30	-4.0	-4.599-4.601-4.600

^a For structures see ref 22.

range segment generation using the same systematic search, the prediction of 20 molecules took only 28 s. Of these compounds, 14 have activity predicted in-range as shown in Figure 3. Clearly, this model has real predictive ability despite the large number of PLS vectors.

Angiotensin Converting Enzyme. Another test set consists of inhibitors of angiotensin-converting enzyme (ACE) as compiled from various sources by Mayer et al.²² with biological activity presented in Table 3 and used as $[g_l, g_u] = \log IC_{50} \pm 10\%$ for 28 active compounds and two inactive. The presence of inactive molecules in the data may provide insight as to molecular features that are disallowed by binding.

Using the notation described above, the DAPPER method found three models at level = [1,1,2,3] with 28 PLS vectors from 37 total descriptors. The method required 29 molecules

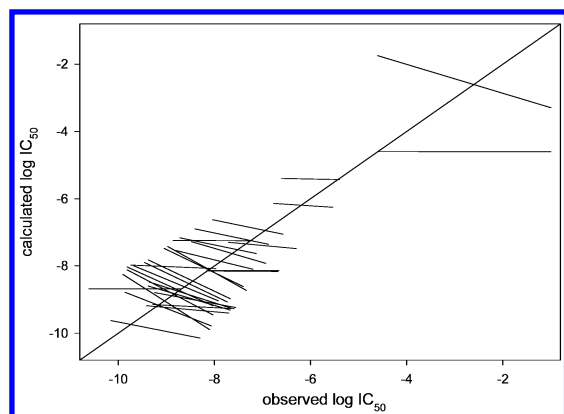


Figure 4. Correlation plot for ACE inhibitors. With more than one equivalent model found, an interval is created in the calculated values as well as the experimental values. The data line segments are therefore drawn as the diagonal of the rectangle representing this interval in both dimensions.

in the training set with only one in the test set and was finished in just over 17 h. The calculated inhibition values are presented in Table 3.

In this case, the plot of observed vs calculated binding presented in Figure 4 has off-horizontal line segments. This occurs because three different DAPPER models were found, and therefore each calculated value then also becomes an interval $[g_{calc,min}, g_{calc,max}]$ and is drawn as a line segment between those values. Several segments may appear horizontal, while others are significantly tilted. This is not surprising as the three models are not required to have identical results, only to have all molecules fit or predicted in-range. Only for molecules where $g_{calc,min} = g_{calc,max}$ are the line segments exactly horizontal.

As a further test of the method, an attempt was made to find equivalent DAPPER models to explain ACE inhibition using randomly shuffled biological data. Normally, DAPPER has the ability to move molecules between the training set and test set, but this feature was eliminated during the randomization test. The composition of the training set was held constant, and the one molecule used as the prediction test in the original run, molecule **21**, remained the lone member of the test set. The inhibition intervals for the molecules in the training set were randomly shuffled, and models were sought at each of 16 different levels of roughly equivalent resolution including the level of the original model.

The inhibition interval for the single test set molecule was used without alteration. In this way, a random model was searched for that would still correctly predict the binding of one single molecule in the correct range. It was thought that predicting only one value correctly would be a simpler task than predicting the multiple molecules presented by the test set of any other system. A correct prediction would indicate that DAPPER is not finding models calculated based only on the experimental data, but that the results likely come from overfitting the data.

The randomized run was performed 20 times with each run taking just under 7 h to search through all given levels. Not one model was found that could accurately predict the single molecule test set, and for all but 22 of the 320 attempts the randomized training set data could not be reasonably fit. This shows that the method is not simply using the large

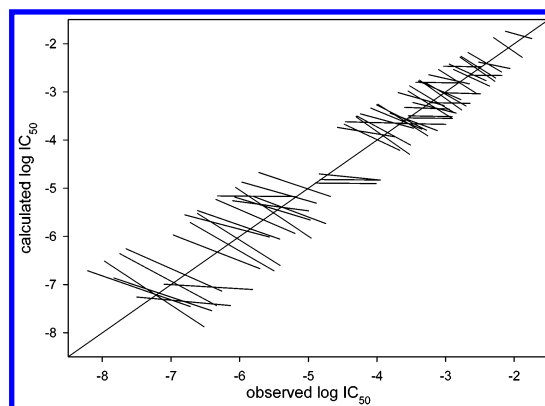


Figure 5. Correlation plot for AChE inhibitors.

number of adjustable parameters to overfit the data but is instead making predictions based on molecular data.

Acetylcholinesterase. A final validation of DAPPER was performed on a set of 59 acetylcholinesterase (AChE) inhibitors.²³ This set was recently the subject of a modified CoMFA calculation, and the data are taken as presented in that study except that one of the original compounds was omitted as it had too many rotatable bonds to be treated in a reasonable time by the systematic range segment search. As with ACE, the experimental data here is presented as the IC_{50} of each compound and is used as the interval $\log IC_{50} \pm 10\%$.

DAPPER found four equivalent models for AChE inhibition at level = [1,2,3,2] (145 descriptors) from 50 PLS vectors and 55 molecules in the training set and 4 in the test set with DAPPER calculated inhibition values shown in Table 4. Again, the plot of observed vs calculated inhibition presented in Figure 5 contains diagonal line segments drawn between $g_{calc,min}$ and $g_{calc,max}$.

It can clearly be seen in Figure 5 that most of the line segments cross the diagonal observed = calculated line near the midpoint of the segment. This could be taken as an indication that the molecule responsible for each of these segments does not factor strongly in the fit of the model. With this in mind, a second version of the AChE data set was run, this time starting with only those molecules for which the fit or predicted inhibition value was more than 5.0% separated from the midpoint of the activity range in the training set. That is, the new training set started with those molecules thought to contribute the strongest information to the model and the remaining molecules comprise the test set.

It was expected that this would lead to the final model being fit with many fewer molecules used in the training set. Interestingly, the single final model in this case was found at level = [1,2,3,2] with 51 PLS vectors from 55 molecules in the training set and 4 in the test set. This result was very similar to the original AChE result with the main difference, albeit slight, in the distribution of molecules between the training and test data sets. Within these and other groups of similar molecules, the exact distribution between training and test sets is likely not critical.

CONCLUSIONS

We have described improvements to our recent program DAPPER and presented further validation of the technique.

Table 4. DAPPER Calculated AChE Inhibition Values^a

no.	log IC ₅₀	<i>g_{calc}</i> (log IC ₅₀)
1	-2.68	-2.551-2.424-2.768-2.427
2	-3.16	-3.288-3.221-3.440-3.356
3	-2.09	-2.286-1.875-1.919-1.883
4	-1.94	-1.800-1.889-1.893-1.744
5	-2.29	-2.383-2.439-2.514-2.452
6	-2.75	-2.481-2.473-2.472-2.479
7	-2.76	-3.019-3.033-3.033-3.040
8	-2.43	-2.189-2.182-2.585-2.224
9	-2.42	-2.659-2.653-2.653-2.662
10	-2.52	-2.364-2.265-2.449-2.713
11	-2.62	-2.880-2.858-2.534-2.590
12	-3.36	-3.286-3.014-3.044-3.204
13	-2.94	-3.228-3.226-3.232-3.239
14	-2.82	-2.841-2.534-3.040-2.605
15	-3.64	-3.798-3.270-3.536-3.373
16	-3.90	-4.106-3.504-3.781-3.924
17	-4.07	-4.214-3.671-3.847-3.970
18	-2.53	-2.281-2.781-2.783-2.674
19	-3.07	-3.378-2.758-3.057-2.900
20	-2.95	-2.659-2.646-2.647-2.832
21	-4.06	-3.649-3.639-3.622-3.632
22	-3.22	-3.513-3.509-3.496-3.515
23	-3.33	-3.659-3.657-3.667-3.673
24	-3.27	-3.329-3.331-3.322-3.359
25	-3.09	-3.158-2.774-3.162-3.082
26	-3.20	-3.426-2.875-2.891-2.884
27	-3.00	-2.805-3.294-2.918-2.854
28	-3.72	-3.520-3.338-3.772-3.403
29	-6.01	-6.425-6.608-5.671-5.527
30	-3.85	-3.469-3.460-3.459-3.753
31	-5.55	-5.267-5.471-5.302-5.263
32	-5.51	-6.011-4.985-5.943-6.036
33	-5.52	-5.661-5.558-5.248-5.187
34	-4.39	-4.827-4.822-4.824-4.831
35	-5.27	-5.708-5.732-5.170-5.354
36	-6.18	-5.562-5.556-6.015-5.565
37	-3.22	-3.540-3.540-3.543-3.550
38	-3.12	-2.812-2.804-2.805-2.812
39	-4.16	-3.748-3.931-3.738-3.915
40	-5.42	-5.308-4.874-4.876-4.883
41	-3.22	-2.987-3.128-3.202-3.549
42	-3.62	-3.259-3.912-3.778-3.319
43	-3.46	-3.444-3.618-3.804-3.557
44	-3.91	-4.266-4.299-3.518-3.551
45	-7.24	-7.260-7.877-6.514-7.736
46	-6.96	-7.136-6.406-6.257-6.265
47	-6.82	-7.363-7.335-7.440-7.264
48	-6.34	-5.972-5.969-6.669-6.158
49	-6.46	-6.993-7.096-7.099-7.107
50	-7.47	-7.457-6.716-6.744-6.751
51	-5.77	-5.942-5.305-5.235-5.261
52	-6.01	-6.058-5.506-5.836-5.464
53	-5.74	-5.173-5.163-5.166-5.170
54	-5.20	-5.171-4.674-5.069-5.166
55	-4.40	-4.819-4.831-4.708-4.746
56	-4.46	-4.900-4.896-4.897-4.905
57	-6.11	-6.720-5.714-6.713-6.137
58	-7.04	-7.120-6.361-7.436-6.639
59	-7.12	-7.138-7.548-7.347-6.865

^a For structures see ref 23.

While the chirality metric presented may not be ideal for all situations, it is a powerful and useful tool as a QSAR descriptor along with or apart from the DAPPER method. It assigns a positive or negative value to each of the two enantiomers of a chiral molecule, and zero to achiral molecules, and moreover, it is simple to calculate and consistent with the concept of atomic property values used in DAPPER. Molecular conformation being an important consideration in any QSAR or molecular modeling technique,

a simple survey identified the standard systematic search of torsion angles to be the most thorough and consistent search method. Unfortunately this technique is also computationally expensive, and its lack of adequate speed illustrates the need for further improvements in this area.

In these examples, generally most of the compounds were required to develop the models, and often the number of PLS vectors employed was not small compared to the number of compounds in the training set. In the context of traditional least-squares fits, this would arouse concerns about overfitting. Here, however, the situation is qualitatively different. Whereas a least-squares fit has the freedom to permit a few large residuals as long as most compounds are well fitted, DAPPER demands that all compounds in the training set have residuals no greater than the chosen limits. This restriction forces the method to use more adjustable parameters and more of the available compounds to achieve the required fit. Within the framework of DAPPER models, the result is one or more models that are the simplest possible while still achieving this rigorous objective. If wider error bars are permitted, DAPPER can find models with fewer parameters using fewer training compounds. This is analogous to linear regression studies where using fewer descriptors gives fits having larger standard deviations. In either approach, predictive power is often taken as the final arbiter of success, and we have demonstrated clear predictive power in the case of 20 DHFR inhibitors.

ACKNOWLEDGMENT

This work was supported by National Institutes of Health Grant GM59097.

REFERENCES AND NOTES

- (1) Makino, S.; Kuntz, I. D. ELECT++: Faster Conformational Search Method for Docking Flexible Molecules Using Molecular Similarity. *J. Comput. Chem.* **1998**, *19*, 1834-1852.
- (2) Chen, X.; Rusinko, A.; Tropsha, A.; Young, S. S. Automated Pharmacophore Identification for Large Chemical Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 887-896.
- (3) Gillet, V. J.; Wild, D. J.; Willett, P.; Bradshaw, J. Similarity and Dissimilarity Methods for Processing Chemical Structure Databases. *Computer J.* **1998**, *41*, 547-558.
- (4) Seri-Levy, A.; West, S.; Richards, W. G. Molecular Similarity, Quantitative Chirality and QSAR for Chiral Drugs. *J. Med. Chem.* **1994**, *37*, 1727-1732.
- (5) Wildman, S. A.; Crippen, G. M. Three-Dimensional Molecular Descriptors and a Novel QSAR Method. *J. Mol. Graphics Modell.* **2002**, *21*, 161-170.
- (6) Crippen, G. M. VRI: 3D QSAR at Variable Resolution. *J. Comput. Chem.* **1999**, *20*, 1577-1585.
- (7) Wildman, S. A.; Crippen, G. M. Evaluation of Ligand Overlap by Atomic Parameters. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 446-450.
- (8) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868-873.
- (9) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity - A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219-3228.
- (10) Moreau, G. Atomic Chirality, a Quantitative Measure of the Chirality of the Environment of an Atom. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 929-938.
- (11) Symon, K. R. *Mechanics*, 2nd ed.; Addison-Wesley: Reading, 1960; p 432.
- (12) Kuz'min, V. E.; Stel'makh, I. B.; Bekker, M. B.; Pozigun, D. V. Quantitative Aspects of Chirality. I. Method of Dissymmetry Function. *J. Phys. Org. Chem.* **1992**, *5*, 295-298.
- (13) Molecular Operating Environment, Chemical Computing Group, Montreal, 2001.

- (14) Cahn, R. S.; Ingold, C.; Prelog, V. Specification of Asymmetric Configuration in Organic Chemistry. *Angew. Chem., Int. Ed. Engl.* **1966**, 5, 385–415.
- (15) Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. A. Conformational Changes of Small Molecules Binding to Proteins. *Bioorg. Med. Chem.* **1995**, 3, 411–428.
- (16) Ferguson, D. M.; Raber, D. J. A New Approach to Probing Conformational Space with Molecular Mechanics – Random Incremental Pulse Search. *J. Am. Chem. Soc.* **1989**, 111, 4371–4378.
- (17) Saunders, M.; Houk, K. N.; Wu, Y. D.; Still, W. C.; Lipton, M.; Chang, G.; Guida, W. C. Conformations of Cycloheptadecane – A Comparison of Methods for Conformational Searching. *J. Am. Chem. Soc.* **1990**, 112, 1419–1427.
- (18) Smellie, A.; Teig, S. L.; Towbin, P. Poling: Promoting Conformational Variation. *J. Comput. Chem.* **1995**, 16, 171–187.
- (19) Hansch, C. A.; Hathaway, B. A.; Guo, Z. R.; Selassie, C. D.; Dietrich, S. W.; Blaney, J. M.; Langridge, R.; Volz, K. W.; Kaufman, B. T. Crystallography, Quantitative Structure–Activity Relationships and Molecular Graphics in a Comparative Analysis of the Inhibition of Dihydrofolate Reductase from Chicken Liver and *Lactobacillus casei* by 4,6-Diamino-1,2-dihydro-2,2-dimethyl-1-(substituted-phenyl)-S-triazines. *J. Med. Chem.* **1984**, 27, 129–143.
- (20) Hansch, C. A.; Li, R. L.; Blaney, J. M.; Langridge, R. Comparison of the Inhibition of *Escherichia coli* and *Lactobacillus casei* Dihydrofolate Reductase by 2,4-Diamino-5-(substituted-benzyl)pyrimidines: Quantitative Structure–Activity Relationships, X-ray Crystallography, and Computer Graphics in Structure–Activity Analysis. *J. Med. Chem.* **1982**, 25, 777–784.
- (21) Selassie, C. D.; Fang, Z.; Li, R.; Hansch, C.; Debnath, G.; Klein, T. L.; Langridge, R.; Kaufman, B. T. On the Structure Selectivity Problem in Drug Design. A Comparative Study of Benzylpyrimidine Inhibition of Vertebrate and Bacterial Dihydrofolate Reductase via Molecular Graphics and Quantitative Structure–Activity Relationships. *J. Med. Chem.* **1989**, 32, 1895–1905.
- (22) Mayer, D.; Naylor, C. B.; Motoc, I.; Marshall, G. R. A Unique Geometry of the Active Site of Angiotensin-Converting Enzyme Consistent with Structure–Activity Studies. *J. Comput.-Aided Mol. Design* **1987**, 1, 3–16.
- (23) Cho, S. J.; Garsia, M. L. S.; Bier, J.; Tropsha, A. Structure-Based Alignment and Comparative Molecular Field Analysis of Acetylcholinesterase Inhibitors. *J. Med. Chem.* **1996**, 39, 5064–5071.

CI0256081