

Three-dimensional molecular descriptors and a novel QSAR method

Scott A. Wildman¹, Gordon M. Crippen*

College of Pharmacy, University of Michigan, 428 Church Street, Ann Arbor, MI 48108, USA

Received 12 November 2001; accepted 15 February 2002

Abstract

A novel set of molecular descriptors suitable for use in quantitative structure–activity relationships and related methods is described. These descriptors are a smooth and interpretable representation of atomic physicochemical property values and intramolecular atom pair distances. Distance atomic physicochemical parameter energy relationships (DAPPER), a novel structure–activity relationship (QSAR) method using these descriptors, is validated on standard datasets.

© 2002 Elsevier Science Inc. All rights reserved.

Keywords: Structure–activity relationships; Three-dimensional QSAR; Atomic physicochemical properties; Computer-aided drug design; Pharmacophore

1. Introduction

Recent improvements in atomic methods for calculation of physicochemical parameters [1] and a measurement of ligand atom overlap in the context of a protein binding site [2], make possible the development of a novel set of molecular descriptors for use in three-dimensional quantitative structure–activity relationship (3D QSAR) studies. This work has the main purpose of prediction of biological activity for unknown, or untested, compounds, and is not intended to be a method for protein–ligand docking [3–5], pharmacophore identification [6,7], de novo molecular design [8,9], or virtual screening [10–12], although use of the descriptors could be extended into any of these areas.

Distance atomic physicochemical parameter energy relationships (DAPPER) will use atomic physicochemical parameters and atom–pair distances at variable resolution to generate a set of 3D descriptors for QSAR which will then be solved using a modified partial least squares (PLS) algorithm. The necessary features are: (1) accurate prediction of activity over a wide range of structures and activity values, including the strict limitation of false positive and false negative results; (2) generation of a deliberately low resolution model in order to avoid overinterpretation of data; (3) smooth changes in the resolution of the model; (4) experimental data treated as intervals rather than specific values; and (5) a graphical representation of the binding

site to allow quick visual inspection of necessary ligand properties capable of explaining existing information. By following these criteria, and avoiding the molecular alignment step common to most existing 3D QSAR methods, DAPPER will escape many of the difficulties commonly associated with current QSAR techniques.

Other QSAR methods meet some of these requirements, and features of these methods were applied directly, some were modified slightly, and others were avoided altogether. DAPPER may therefore seem like an unwieldy mixture of techniques and ideas, but it is firmly based on previous QSAR techniques by Crippen [13], most recently, and by others. This work was completed using the molecular operating environment (MOE) [14].

2. Methodology

2.1. Dataset preparation

The first step in the DAPPER technique is to divide the data into two disjoint subsets: a training set from which the QSAR model will be constructed, and a test set which will be used to accept or reject proposed models. At first, the training set is comprised of the three tightest-binding (or otherwise highest activity) molecules, and the test set is comprised of all molecules not part of the training 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. If no model can be fit at this resolution, the level is gradually increased until a model can be found

* Corresponding author.

E-mail address: gcrippen@umich.edu (G.M. Crippen).

¹ Present address: Pfizer Global Research and Development, Ann Arbor, MI, USA.

to fit the data. When a model is found that fits 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 training set and accurately predict the biological activities in the test set, the model is accepted, and the program ends. If, however, none of the models predicts accurately at several successive levels, the molecule in the test set with the greatest error in prediction is moved from the test set into the training set, and the data subsets remain disjoint. Additionally, if a predefined high level is reached with a given training set and no accepted model is found, again the worst predicted molecule is moved from the test set into the training set, and in this case the level is reset to the starting low resolution.

On the other hand, if a molecule in the training set is predicted near the midpoint of the activity range for several models, and none of those models passes the prediction test, then that molecule may be returned from the training set to the test set and the subsets remain disjoint. It is assumed that this easily fit molecule is not adding any significant constraint to the model and therefore is unnecessary for training. If this assumption turns out to be incorrect and the given molecule becomes the worst predicted in the test set, it will be returned to the training set by the standard criteria.

Atomic contributions to octanol–water partition coefficient ($\log P$) and molar refractivity as calculated by the SLOGP/SMR method along with Gasteiger–Marsili partial charges [15] are assigned to each atom of each molecule. This provides a description of each atom in terms of hydrophobicity, steric bulk and electrostatics, and along with intramolecular atom pair distances, these data will be used to construct the DAPPER molecular descriptors.

2.2. Conformational flexibility

DAPPER incorporates 3D molecular information by considering the Cartesian distances between all intramolecular atom pairs. These distances are defined by molecular conformation, and therefore correct treatment of molecular flexibility is a necessary part of DAPPER, as it is for any accurate 3D molecular modeling method. Such treatment is often neglected for algorithmic ease, but also commonly because complete conformation searching, using any of several available methods, can be considerably time consuming. The QSAR methods that do include some treatment of molecular flexibility usually perform a conformation search separately from the rest of the method, storing a discrete list of conformations, usually sorted by energy, for use in the subsequent technique. However, considering conformations from a discrete list is not an accurate representation, as molecular conformation space is continuous. That is, if a molecule can adopt a conformation with the distance between atoms i and j of $d_{ij} = m$, and it can adopt another conformation with $d_{ij} = n$, then the molecule can also find conformations to produce every distance between m and n although admittedly, some of these conformations and distances may be infeasible when including energy considerations.

A better treatment of molecular flexibility would incorporate all implied conformations as well as those listed. A reasonable way to accomplish this is to represent a set of conformations as a distance range, where each interatomic distance is defined by the minimum and maximum possible distances over all conformations [16]. Just as a single conformation may be defined by a list of all interatomic distances, all of conformation space may be represented as a list of all interatomic distance ranges. Keeping these endpoints of each atom pair distance range allows for the occurrence of any distance within that range, not just those from conformations present in the discrete list. While there is no assurance that the active conformation of a molecule (which will vary with different biological systems) is actually present in a discrete list, the distance range technique guarantees that the active conformation (for any biological system) is represented if the conformational search pushes each atom pair to its distance extremes, subject to the constraint of reasonable energy.

A more difficult issue arises from correlations between distances when considering more than one atom pair simultaneously. While the entire distance range is possible for any single atom pair, multiple pairs may have, and usually do have, correlations between these distances. A good example would involve two atoms in a rigid substructure, say *ortho* substituents on an aromatic ring. These atoms will always be close to each other, and therefore their distances to other atoms are highly correlated. When measuring distances of these atoms to a remote atom attached by a flexible linker, it may be possible to find some conformations with short distances, as shown in Fig. 1a, and some with long distances, Fig. 1b. While some variation in distance may be allowed (as indicated by the boxed arrow heads in the figure) the xy and xz distances are correlated to the extent that there can not be any conformation with a long xy distance and a short xz distance.

DAPPER avoids this situation by using several small subsets, or segments, of the distance ranges rather than an entire range for each atom pair. In Fig. 1c, the overall range is divided into a few segments labeled r_1 – r_5 , and within each segment all distances are possible. Each of the range segments creates a continuous subset of conformation space which still allows for local flexibility, and since the segments come from adjacent conformation space, all feasible distances are represented in a still continuous manner. This method avoids the potential problem described by the triangle inequality and therefore prevents a QSAR model from being based on a conformation with a long xy distance and short xz distance.

A more abstract representation is shown in Fig. 2a, where if only the minimum and maximum distances are used to generate the ranges, all the distances would be included even though much of this area would not represent feasible combinations of distances. Only conformations in the small ‘feasible ranges’ area can actually occur and therefore only this area should be included in any representation of distance.

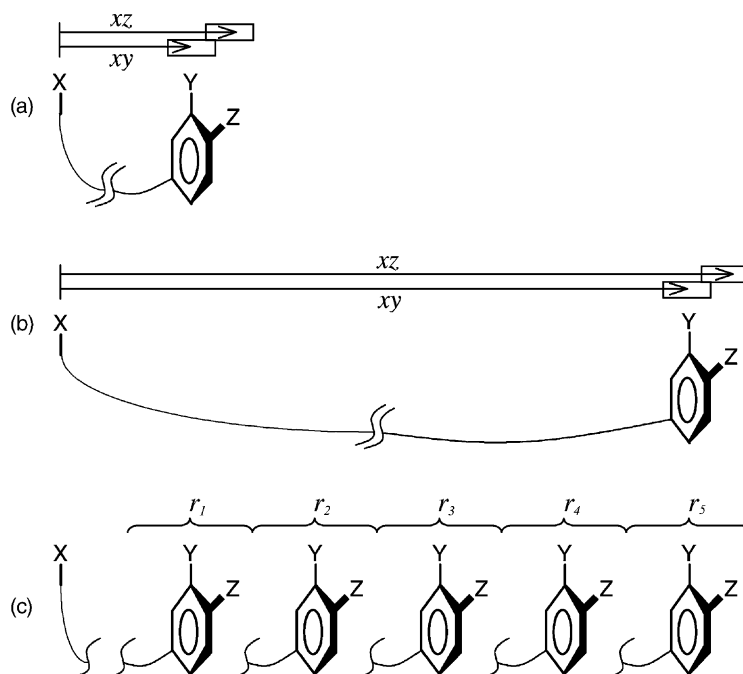


Fig. 1. Distance ranges and distance range segments. For a molecule with rigid structure between atoms Y and Z and an atom X connected by a flexible linker, the xy and xz distances must both be short (a) or long (b). It is not feasible to have a short xz and long xy simultaneously. As shown in (c) the range can be divided into smaller segments, labeled r_1 – r_5 .

Fig. 2b, shows the treatment of distance range segments in which each of 22 smaller squares represents a distance range segment, allowing for some specific range in each distance, and covering the feasible range area completely. There is some overlap with infeasible range space, shown by dark diagonal stripes, which will depend on the set range segment size, but this error is small compared to the entire infeasible space. If the range segment boxes are chosen to be large, fewer segments are needed to cover the feasible space, but the overlap with the infeasible region will be larger. On the other hand, if a smaller box size is chosen, the error is decreased, but the number of range segments needed to cover feasible space rises, thus complicating the representation of the system and the remaining calculations. It is also important to note that Fig. 2 shows a simple case of two correlated distances, while molecules having n atoms have $n(n-1)/2$ pairs of atoms, and therefore the true plot of this feature would have $n(n-1)/2$ dimensions.

DAPPER implements the range segment search using a modification of this MOE [14] systematic search algorithm. This process generates molecular conformations by systematically rotating bonds in a molecule by small increments, for all rotatable (single, non-terminal) bonds, with ring flips included for aliphatic rings. Once the search is complete, conformations having internal energy within 7 kcal of that of the lowest energy conformation found have all atom pair distances measured and considered for range segments. This distance data is used to generate a list of range segments, where the width limit of each segment is set at 6.25 Å. When

comparing the atom pair distances from two conformation search steps, if the difference in any distance between those steps is longer than this limit, other existing segments are checked, and if necessary, a new range segment is started.

The process will commonly generate tens of range segments, each of which may be thought of as a different conformation of a molecule, where each “conformation” really is a representation of a small but continuous portion of conformation space. In DAPPER, these range segment “conformations” will each be treated as separate representations of the same molecule, or pseudo-molecules, and a separate descriptor list will be generated for each. The dataset can then be represented as follows:

$$\begin{aligned} m_1 &\Rightarrow r_1, r_2, r_3 \\ m_2 &\Rightarrow r_1 \\ m_3 &\Rightarrow r_1, r_2, r_3, r_4 \\ m_4 &\Rightarrow r_1, r_2, r_3 \end{aligned}$$

where each m_i is a molecule in the dataset and each r_j is a range segment of that molecule. DAPPER will require that at least one range segment for each molecule bind with sufficient calculated affinity, and none bind too tightly. This requirement is akin to that made by actual ligand binding, where one conformation will bind to provide the observed binding constant and any other conformations that bind must do so with lower affinity.

As this search is based on a systematic conformation search, it can be rather time consuming. Only very small, rigid molecules go through the process quickly. Therefore,

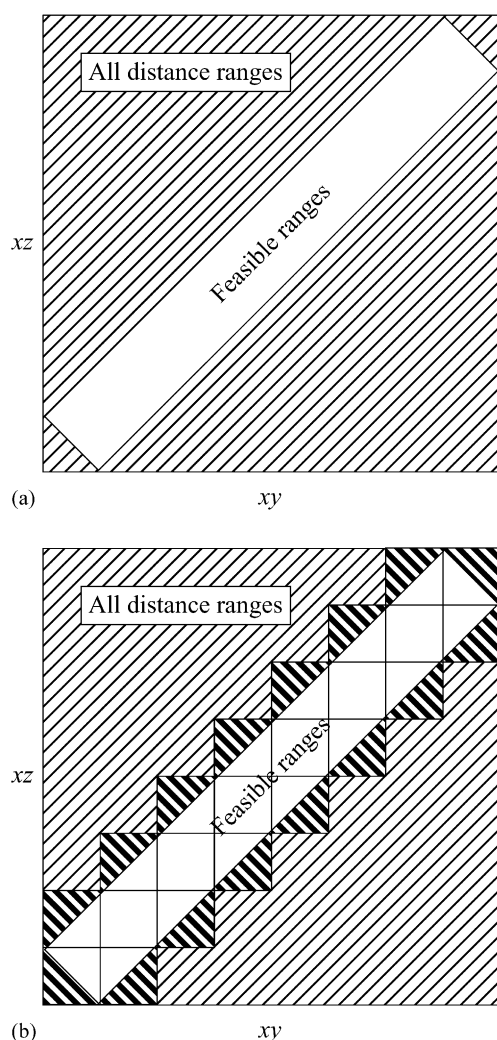


Fig. 2. Range segment coverage of feasible range space. For correlated distances, xy and xz , the overall distance range is large, but only a small region of this area is feasible. Each of the 22 smaller squares in (b) represents a distance range segment. These completely cover the feasible region and overlap with a minimum of the infeasible area, shown as dark diagonal stripes.

in DAPPER, this range segment search is completed once, and the endpoints of all atom pair distance ranges for all segments are stored. This would also allow for simple modification of the range segment portion of the method without changing the remainder of the DAPPER program.

2.3. Descriptor generation

For each molecule, the atomic property values and atom pair distance ranges can be thought of as a distribution of data in property space. One way to represent this sort of distribution would be a histogram in each of the property dimensions. Using histograms, each property (including distance) can be divided into class intervals of the property space. These intervals, also called bins, simply categorize

the data placed in them. For instance, a two bin histogram for SLOGP would divide the data into intervals of low SLOGP and high SLOGP, the exact boundaries of which can be defined arbitrarily. Adding more bins over the same property scale effectively increases the detail of the histogram representation as the intervals become necessarily narrower when more are used in a constant space. The usual situation is to define a total possible range of data values and a number of bins, the widths of which are then simply the total range divided by the number of bins.

From the standpoint of DAPPER, the standard histogram bins have one major drawback in that bin walls are hard boundaries. This means small changes in data value in the center of a bin will not result in a change of bin occupancy, but if there is a small change in value at or near a bin boundary, the data point may move into the neighboring bin. The hard edge feature of histograms will cause DAPPER to potentially treat very similar molecules as being very different. For this reason it is necessary to find a suitable smooth-wall histogram.

One suggestion to this end would be to have overlapping bins. In this scenario, all data points that fall near a boundary would be in a region where the higher bin had started but the lower bin had not yet ended, and therefore, that data point could be thought of as being in both bins. Careful analysis shows that this is not a fair treatment, as this situation would be indistinguishable from the case of two separate points, one in each bin. If, alternately, a special bin designation is used for the overlap region, the previous hard boundary situation is not relieved, but simply moved to a different location on the property scale.

The more sensible solution is to use a smooth function representation of a histogram in which small changes in data value can only result in small changes of function value. In DAPPER, Gaussian functions are used for this purpose as they have a simple form and position and width are easily controlled. Each histogram bin will now become a Gaussian, as shown in Fig. 3 for instance, where three Gaussians ($f_1(x)$, $f_2(x)$, $f_3(x)$) will be used instead of three histogram bins.

The parameters of each Gaussian are set so that it reaches a maximum of 1.0 at the midpoint of the corresponding histogram bin, the value of $f = 0.7788$ at the limits of its bin, and of course it approaches zero far away from the bin. If we have a set of observations x_1, \dots, x_n , let the value of each "bin" j be $\max_i f_j(x_i)$. This way, small changes in the x_i give only small changes in the Gaussian values, and the result is much the same when large or small numbers of observation are used.

Fig. 3 shows the case of one property dimension (on the x -axis), but the situation in DAPPER is more complicated. Each of the data types is evaluated on Gaussians in a different dimension. For each atom pair, there is a distance range minimum and maximum, SLOGP value of each atom, SMR value of each atom, and partial charge of each atom, for a total of eight property dimensions. The final form of the

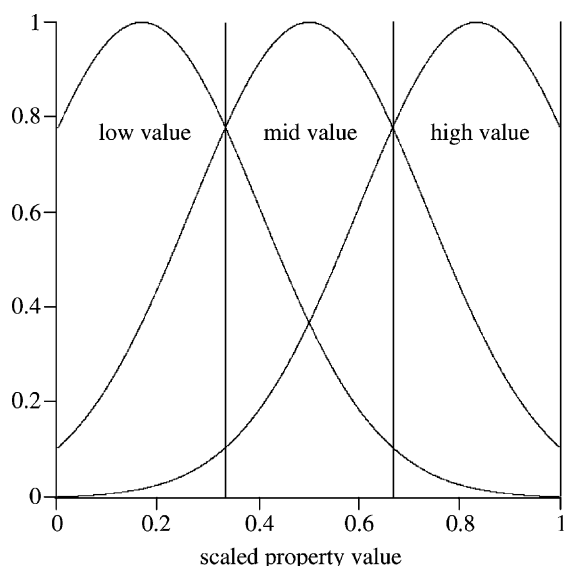


Fig. 3. Gaussian representation of a three-bin histogram. Three Gaussians are positioned over the bins of a three-value histogram. The Gaussians can then be thought of as representing low, middle and high property values, respectively. The Gaussians provide a smooth representation of the property space while maintaining the categorization qualities of the histogram.

eight-dimensional Gaussian is:

$$f(x) = \exp \left[- \sum_{p=1}^8 \alpha_p (x_p - x_{0,p})^2 \right]$$

where the sum is over each (p) property and distance which make up each of the eight dimensions and each $\alpha_p = n_p^2$, the square of the number of Gaussians for that property dimension. This forms, in essence, a grid of Gaussians in eight property dimensions. Each data point is still evaluated on each Gaussian, but most result in near-zero values which do not affect the model as only the maximum observed value of each Gaussian becomes a QSAR descriptor.

In order to maintain consistent values and make for easier generation of the Gaussian grid, each property and distance is rescaled to $[0, 1]$ against the maximum and minimum possible property values and an arbitrary distance maximum. Atomic hydrophobicities (h), as measured by SLOGP, range from -2.996 to $+0.8857$, and are scaled as $h' = (h + 2.996)/(0.8857 + 2.996)$, SMR values (r) are scaled against the maximum value of 14.02 with a minimum of 0.0 as $r' = r/14.02$, partial charges (q) are scaled using $q' = (\min[1, \max[-1, q]] + 1)/2$, and distances are scaled against a presumed maximum distance of 25 \AA , using $d' = d/25.0$.

Since the descriptor list is written from the maximum observed values of each Gaussian in the eight-dimensional grid, it is possible to identify which section of the grid, and therefore which combination of properties is responsible for each of the descriptor values. When a final model is found, the values in the coefficients of the fit can be used as an

indication of which physical properties and which distances are most important to the QSAR model, and a graphical model of this information may be constructed.

2.4. Level of resolution

Over the same property scale, a two bin histogram (low and high property value) is of lower resolution than a three-bin version with low, medium and high values. Each of these three bins will be narrower than those in the two bin version and therefore as the number of bins is increased, the resolution, or degree of detail, also increases. The same situation occurs with the Gaussian representation described in the previous section. The number of Gaussians used to represent each property dimension can be varied, and the level of resolution changes too.

One might expect that a high resolution model could better predict the features necessary for the new drug by including as much detail as possible, however this is not necessarily the case. As a more detailed description is produced, at some point it becomes quite likely that the model is overinterpreting the data. Since a method can only use the molecular information present in the dataset, the resulting model, if too detailed, will become biased towards these data. While this might result in a good model fit, the predictive ability of the model is reduced, particularly for molecules with significantly diverse structure. Instead, the model should be built with as low resolution information as possible. This way, the model will still have acceptable fit, as determined by separate criteria, but without sacrificing predictive ability. It is possible that this approach will result in a model with such low resolution that many compounds are identified as having activity, but this is seen as a better alternative than potentially rejecting a molecule that would indeed be active due to a biased high resolution model.

The DAPPER method attempts to produce a deliberately low resolution model by starting at a very low resolution and gradually increasing resolution until an acceptable model is found. To start, only one Gaussian in each of the eight property dimensions is used and the resulting model can provide only very limited information about the physical properties and distances present in the dataset, and should therefore not be expected to produce an accurate and predictive model. When this model fails as expected, the resolution of the data representation is increased slightly, and the resulting new model is attempted. This process is continued for subsequent levels, with each model being subjected to the testing criteria until an acceptable model is found. If a given model can not both accurately fit the training set and predict the test set, the model is rejected, and the level of resolution is increased. As this process will stop once a model is found that passes these criteria, the resolution must be increased in order of lowest resolution first.

In DAPPER, increasing the resolution corresponds to using more Gaussians in one or more property dimensions. The level could, in theory, be varied in all dimensions

independently, but in order to have a more consistent view, the detail will change in pairs of dimensions. That is, the resolution in both of the distance dimensions will vary together, as will the two SLOGP dimensions, the two SMR dimensions, and the two partial charge dimensions. Thus, the level can be described as a simple vector, [D, H, R, Q], with a single element each for distance, hydrophobicity (SLOGP), refractivity (SMR), and partial charge respectively. Using this notation, [2, 1, 3, 2] would represent two Gaussians in each of the distance dimensions, one in each SLOGP dimension, three in each SMR dimension, and two in each charge dimension.

It is the level of resolution that determines the number of descriptors in DAPPER. The number of descriptors from the Gaussians is the product of the squares of the entries of the level vector, as each of those entries represents the number of Gaussians in two property dimensions. To each descriptor line, a constant equal to 1 is added, and chirality descriptors [17] may also be included.

2.5. Data fitting

Traditional QSAR methods generate a matrix of descriptors C and have a vector of observed biological activities y , so that the m th row of C are the descriptors of molecule m , and y_m is the activity for molecule m . The object is then to find a model vector v such that Cv approximates y in a least squares sense, that is, minimizes $\|Cv - y\|^2$ with respect to v .

At all but the lowest levels of resolution, a large number of descriptors result from the Gaussian method of generation. As there are likely far fewer observations present in the data, it becomes necessary to reduce the number of descriptors using PLS as described by Wold and coworkers [18]. However, the resulting PLS vectors are not fit in the traditional manner, but instead the method employed follows our recent adaptation [13].

In DAPPER, there are two special circumstances that must be considered. First, for each molecule m there can be multiple lines of descriptors resulting from multiple distance range segments. Second, in order to treat both precise and imprecise biological data, error bars are placed on each single data value. Since this approach is quite different from the standard single point value representation of activity data, different techniques must be used to fit the model.

This biological data interval is presented as $[g_{m,l}, g_{m,u}]$ where $g_{m,l}$ and $g_{m,u}$ are the lower and upper limits on the error bar interval for each molecule, respectively. In this description, the calculated activity for a given range segment r of a given molecule m , denoted by $g_{m,r,calc} = c_{m,r}v$ is considered superoptimal if $g_{m,r,calc} < g_{m,l}$, suboptimal if $g_{m,r,calc} > g_{m,u}$, and otherwise in-range. In other words, lower values indicate tighter binding or greater activity.

With interval biological data and several lines of descriptors for each molecule, a natural thought is to require that for each molecule at least one range segment must have at least one $g_{m,r,calc}$ in-range on the interval, and none may be

superoptimal. This is akin to actual ligand binding, where one conformation will bind to provide the observed binding constant, and any other conformations that bind must do so with lower affinity. In DAPPER, this is accomplished by adjusting v to minimize the penalty function:

$$F(v) = \sum \begin{cases} \sum_{\text{super } r} (c_{m,r}v - g_{m,l})^2, & \text{any superoptimal} \\ \frac{n_m^2}{(\sum_{r=1}^{n_m} (c_{m,r}v - g_{m,u})^{-1})^2}, & \text{all suboptimal} \\ 0, & \text{otherwise} \end{cases}$$

where $F \geq 0$. In the case where all n_m range segments of molecule m are suboptimal, $F \rightarrow 0$ as any one or more $c_{m,r}v \rightarrow g_{m,u}$, yet any single superoptimal range segment will result in $F > 0$. Optimizing to $F = 0$ results in the situation where all the molecules have at least one $g_{m,r,calc}$ in-range and none superoptimal. Accomplishing this with the training set is the requirement for sufficient fit of the data. Unlike least squares fitting of a training set, each model either is in full agreement with the given binding intervals, or it is rejected altogether.

Since this method fits to an interval of experimental activity, rather than a standard least squares fit to single values, the solution v is not uniquely determined. Therefore, it is also necessary to carry out a simple random search for any perturbation w such that $F(v + w) = 0$ and include these solutions as long as they are sufficiently different. Each entry in this list of fitting solutions is subjected to the test set prediction criteria, and only those solutions that pass are accepted as the final model.

2.6. Graphical representation

It is possible to generate a graphical representation of the features necessary for activity based on the final DAPPER model, resulting in a picture which can be useful to explain the existing data. The picture is constructed by using the QSAR model coefficients to identify the small number of Gaussians that make the greatest magnitude (positive or negative) contribution to activity for a given molecule, usually the most active, in a given model. Therefore, systems with multiple DAPPER models will have more than one such picture. It is also necessary to identify which of the distance range segments provides the greatest activity for the given molecule. Since each segment was treated on a separate descriptor line, it is necessary to reconstruct the correct line of descriptor values only. Once the most important Gaussians are found, the atom pair responsible for the maximum value of each of those Gaussians is identified from the descriptor values and is highlighted in the molecule. Physical property information may be coded by color and distance range tolerance may be indicated by drawing appropriately large or small atoms for each important atom pair.

2.7. Summary

Following separation of the data into subsets for training and testing and completion of the distance range segment search on all molecules, the full set of descriptors is determined for each range segment of each molecule. This is accomplished by evaluating all distance range and atomic property data on a grid of Gaussians in eight property dimensions where the number of Gaussians in each dimension is described by the level vector, and taking the maximum observed value of each Gaussian as the descriptor value. The training set data is fit to the biological activity data, presented as intervals, using a modified PLS, and a penalty function, $F(\mathbf{v})$, is minimized to create a model for that level.

This model is then used to predict the activity of the molecules in the test set, and any solution for which all molecules pass the prediction test becomes a valid final model. If no proposed solution passes the acceptance criteria, the model is rejected, the level is incremented and a new model is generated.

3. Results and discussion

3.1. Artificial examples

Each DAPPER solution can be comprised of several different models at a given level of resolution. Therefore, each set of results needs to define the level and provide the PLS adjusted coefficients for each model. Other values of interest, although not strictly necessary, are the number of molecules used in the training and test sets, the number of PLS vectors needed for an acceptable fit, and the time needed to calculate the model. While looking at the coefficient values themselves is not terribly descriptive, these values are the main result of the method and are used to predict biological activity data for previously unknown molecules.

Before considering standard test data, it is useful to illustrate DAPPER's performance in differentiating simple pairs and small sets of molecules that differ in various ways. Some of these examples were designed to identify differences in atomic property values, others for position or distance, and others for chirality. In each case, some molecule(s) were arbitrarily said to be active and others inactive. As these are only simple tests, the coefficients that comprise the models will not be presented in favor of discussion of the qualitative features of the solution.

Four tests were used to discriminate atom property values, the descriptions and results of which are presented in Table 1, tests 1–4. Methane and ethane are both comprised of only C1 and H1 SLOGP atom types, but they differ in that ethane contains a C1–C1 atom pair whereas methane contains only C1–H1 and H1–H1 pairs. This property difference is sufficient to distinguish the molecules even at level = [1, 1, 1, 1], one Gaussian in each of the property dimensions. Likewise, when comparing methane and Cl₂,

the only differences lie in property values, and these examples are also able to be fit at level = [1, 1, 1, 1]. A more complicated test of a series of mono- and dichloronaphthalenes is similar to the methane/ethane test in that the disubstituted molecules have a Cl–Cl atom pair not present in the monochloronaphthalenes. This test proved to be more difficult to solve, but DAPPER still found an acceptable model at level = [1, 2, 3, 3] and only took 2 min to do so.

Tests 5–7 were used to differentiate small differences in distance, and also passed easily. In *cis*-dichloroethylene, the Cl atoms are separated by 3.2 Å, while in the *trans* isomer they are 4.3 Å apart. This small difference in distances translates through the Gaussians into sufficient difference in QSAR descriptors to be fit at level = [1, 1, 1, 1]. A similar test with *ortho*- and *meta*-substituted aromatics finds eleven DAPPER models at level = [1, 1, 1, 3], correctly differentiating between long and short distances of the pair of atoms with the most negative partial charges, the Cl–Cl pairs. A pair of differently substituted chloro-cubanes produces a similar result at level = [1, 1, 1, 4]. Here 1,3,5,7-tetrachlorocubane has a tetrahedral arrangement of chlorines and therefore all Cl–Cl atom pairs have the same distance (5.0 Å), while the 1,4,5,8-substitution pattern results in a planar rectangle of chlorines with three defining pair distances (3.6, 5.1 and 6.3 Å for two sides and a diagonal respectively). This model correctly identifies that a low charge (i.e. rather negative unscaled charge) pair, such as Cl–Cl, at longer distance is detrimental to activity.

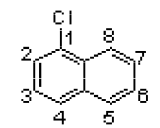
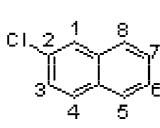
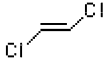
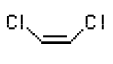
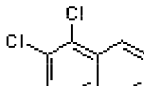
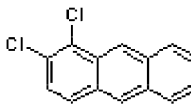
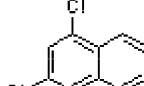
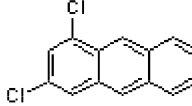
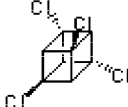
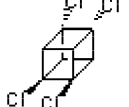
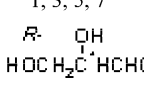
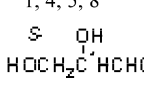


For chirality, tests 8 and 9, used to distinguish stereoisomers, were completed on *R*- and *S*-glyceraldehyde and on *R*- and *S*-fluorochlorobromomethane. Both of these examples were easily solved at level = [3, 1, 1, 1], the first level to include a chirality (χ) value. This is the expected solution as each pair of molecules will be identical in property values and distance range segments, and for both examples, only the constant and chirality descriptors are used to fit the model. In these examples, the non-chirality descriptor values are equal for both the active and inactive molecules and thus, the Gaussian-derived descriptor coefficients are zero.

3.2. CBG and TBG

The standard corticosteroid binding globulin (CBG) and testosterone binding globulin (TBG) datasets were used as two of several real biological data benchmarks. The commonly used dataset [19] is known to contain errors, and while several other sources claim to have corrected the previous errors, their corrections were often either incomplete or additional errors were introduced. The correct steroid structures and binding data are reproduced from the original studies in a recent review [20] and these correct data are used in this work.

Even using the correct structures and values, there are several ways to approach the steroid data, as some authors use all 31 CBG data values to produce their model, and others construct the model using only the first 21 compounds,

Table 1
Artificial examples

| Test | Active | Inactive | Level | Models | Result |
|------|--|--|--------------|--------|-----------------------------------|
| 1 | CH ₄ | CH ₃ CH ₃ | [1, 1, 1, 1] | 1 | C–C pair unfavorable |
| 2 | Cl ₂ | CH ₄ | [1, 1, 1, 1] | 2 | C–H pair unfavorable |
| 3 | CH ₃ Br | CH ₃ Cl | [1, 1, 1, 1] | 1 | Cl–C pair unfavorable |
| 4 | Dichloronaphthalenes 1,2- 1,7- 1,3- 1,8- 1,4- 2,3- 1,5- 2,6- 1,6- 2,7- |   | [1, 2, 3, 3] | 3 | Cl–Cl pair favorable |
| 5 |  |  | [1, 1, 1, 1] | 2 | Longer distance favorable |
| 6 |   |   | [1, 1, 1, 3] | 11 | Longer Cl–Cl distance unfavorable |
| 7 |  1, 3, 5, 7 |  1, 4, 5, 8 | [1, 1, 1, 4] | 1 | Longer Cl–Cl distance unfavorable |
| 8 |  R OH HOCH ₂ C HCHO |  S OH HOCH ₂ C HCHO | [3, 1, 1, 1] | 1 | Chirality only |
| 9 |  Cl H Br F |  Cl H Br F | [3, 1, 1, 1] | 2 | Chirality only |

reserving the last 10 for separate prediction tests. Results from both approaches are presented here. All 21 molecules in the TBG dataset were used to generate the DAPPER model. Biological data in all steroid data was used as $-\log K \pm 10\%$, although tests indicate the exact magnitude of the error bars is not critical.

For both the CBG and TBG datasets, the complete systematic search normally used to produce the distance range segments is prohibitively time consuming. Therefore, a modified range search technique was employed, where the conformations used to derive the range segments were generated using a stochastic search in which rotatable bonds are perturbed by some random increment followed by energy minimization [21] available in MOE version 2001.01. It is thus possible that the full systematic search would result in different range segments and therefore different DAPPER models.

Using all 31 molecules in DAPPER (CBG31), the CBG dataset was fit with 17 molecules in the training set and 14 in the test set at level = [2, 1, 1, 3] to produce two models that passed the penalty function acceptance criteria. This

level results in 38 descriptors which were reduced to 10 PLS vectors. The calculated binding values for both models are presented in Table 2 under the CBG31 heading. The stochastic distance range search took just under 8 h, but the remainder of the method was finished in just under 1 h on a SGI R5000 300 MHz processor (O2 model).

With the alternative 21 molecule training/test set used by DAPPER and the remaining 10 compounds reserved for true prediction (CBG21), DAPPER identified one model at level = [1, 2, 2, 1] (17 descriptors) using 14 molecules in the training set, seven in the test set, and 11 PLS vectors. Of the 10 prediction compounds, six were predicted in-range with the remaining four all predicted to bind more tightly than the $g_{m,l}$ (considered superoptimal prediction), as can be seen in Table 2 under the column labeled CBG21. These results do not produce a high level of confidence in the method, but it should be noted that the prediction set contains several molecules known to be difficult to predict [22]. The range segment search did not need to be repeated as the segment data were stored at the end of the original search, and the

Table 2
CBG and TBG experimental data and DAPPER results

| Steroid | CBG ^a | TBG ^a | CBG31 ^c | CBG21 ^d | TBG21 |
|---------|-------------------------------|-------------------|--------------------|---------------------|--------|
| 1 | [−6.907, −5.651] | [−5.854, −4.790] | −6.907 −6.907 | −6.908 | −5.821 |
| 2 | [−5.500, −4.500] ^b | [−10.025, −8.203] | −5.500 −5.500 | −5.378 | −8.692 |
| 3 | [−5.500, −4.500] ^b | [−10.094, −8.258] | −5.050 −5.017 | −4.867 | −8.290 |
| 4 | [−6.339, −5.187] | [−8.208, −6.716] | −6.337 −6.339 | −6.205 | −7.275 |
| 5 | [−6.174, −5.052] | [−7.861, −6.431] | −5.590 −5.592 | −5.673 | −6.924 |
| 6 | [−8.669, −7.093] | [−6.976, −5.708] | −7.658 −7.649 | −7.112 | −6.538 |
| 7 | [−8.669, −7.093] | [−6.824, −5.584] | −7.838 −7.828 | −7.094 | −6.510 |
| 8 | [−7.581, −6.203] | [−7.074, −5.788] | −7.108 −7.103 | −7.196 | −6.965 |
| 9 | [−5.500, −4.500] ^b | [−8.601, −7.037] | −4.925 −4.891 | −5.146 | −7.373 |
| 10 | [−8.418, −6.888] | [−8.118, −6.642] | −7.859 −7.835 | −7.031 | −6.955 |
| 11 | [−8.669, −7.093] | [−7.924, −6.484] | −7.096 −7.093 | −7.198 | −7.265 |
| 12 | [−6.511, −5.327] | [−10.714, −8.766] | −5.327 −5.327 | −5.470 | −8.766 |
| 13 | [−5.500, −4.500] ^b | [−9.715, −7.949] | −5.270 −4.732 | −4.583 | −8.375 |
| 14 | [−5.500, −4.500] ^b | [−7.296, −5.970] | −4.787 −4.828 | −4.590 | −7.296 |
| 15 | [−5.500, −4.500] ^b | [−8.994, −7.358] | −5.091 −4.813 | −4.501 | −7.751 |
| 16 | [−5.781, −4.730] | [−6.761, −5.531] | −5.748 −5.752 | −5.748 | −6.708 |
| 17 | [−5.781, −4.730] | [−7.861, −6.431] | −5.100 −5.066 | −5.262 | −6.434 |
| 18 | [−5.500, −4.500] ^b | [−6.998, −5.726] | −5.212 −5.179 | −5.313 | −6.235 |
| 19 | [−8.118, −6.642] | [−7.638, −6.250] | −7.055 −6.815 | −8.047 | −6.216 |
| 20 | [−8.514, −6.966] | [−7.638, −6.250] | −8.023 −7.849 | −7.043 | −6.297 |
| 21 | [−7.396, −6.052] | [−10.124, −8.284] | −6.192 −6.185 | −7.298 | −8.307 |
| 22 | [−8.265, −6.763] | | −7.386 −7.340 | −6.955 ^e | |
| 23 | [−8.309, −6.799] | | −7.344 −7.343 | −7.220 ^e | |
| 24 | [−7.457, −6.101] | | −7.191 −6.348 | −7.342 ^e | |
| 25 | [−7.921, −6.481] | | −6.815 −6.811 | −7.218 ^e | |
| 26 | [−6.757, −5.529] | | −6.031 −6.023 | −7.181 ^e | |
| 27 | [−6.872, −5.622] | | −6.764 −6.758 | −7.233 ^e | |
| 28 | [−7.839, −6.413] | | −7.244 −6.963 | −7.106 ^e | |
| 29 | [−7.565, −6.189] | | −6.904 −6.789 | −9.233 ^e | |
| 30 | [−8.437, −6.903] | | −6.905 −6.903 | −6.987 ^e | |
| 31 | [−6.376, −5.126] | | −6.369 −6.376 | −7.166 ^e | |

^a Experimental values given as [$g_{m,l}$, $g_{m,u}$] from: $-\log K \pm 10\%$.

^b Binding affinity reported as $K > 1 \times 10^{-5}$ M.

^c All 31 steroids used in model generation. Values are given for each equivalent model.

^d Only steroids 1–21 used in model generation. Steroids 22–31 are true prediction.

^e Predicted values based on CBG21 model.

remaining DAPPER run took only 8 min on the same processor. This large difference in time is the result of DAPPER identifying a lower resolution solution, likely due to several difficult molecules being in the prediction set rather than the training or test sets, as the time of calculation increases dramatically with model resolution level and number of molecules in each set.

The TBG dataset (TBG21) was analyzed by DAPPER with all 21 molecules in the training and test sets, and one model was identified at level = [1, 2, 3, 1] using 37 descriptors, and 14 PLS vectors with 16 molecules in the training set and five in the test set in 12 min. The calculated binding values are given in Table 2. Again, the stochastic range segment search would not need to be repeated as the original range segments remain valid.

It is important to note that the CBG and TBG sets are likely not the best data to use for comparison of methods. Many versions of the data exist, with varying numbers of structural errors and often data values in disagreement with the original works. It was also illustrated by Kubinyi [22]

that the 10-molecule prediction set of CBG21 contains several structural features not present in the 21 training compounds. Additionally, the original data values for 7 of the 21 training compounds are measured as $K > 1 \times 10^{-5}$ M. No accurate binding affinities for these molecules could be determined, and it can therefore be argued that one-third of the training set be listed as ‘inactive’ or simply not used in the calculation. Considering this information, accurate prediction should not be expected.

4. Conclusions

The novel method for the development of 3D QSAR, DAPPER along with validation has been described in detail here, and implemented in the MOE [14]. The computer programs used in this work are available from the corresponding author on request.

Each of the listed requirements of the method has been met, and accurate prediction of activity is required for each

model through the use of separate training and test sets. The final QSAR model, or set of equivalent models, is the lowest resolution model at which the training data can be accurately fit and the test data can be accurately predicted. Additionally, the resolution of the model can be changed in a relatively smooth manner simply by incrementing the level. This technique is a recent contribution to the field of QSAR, and combined with the Gaussian-based DAPPER descriptors, represents a novel approach to 3D QSAR.

Acknowledgements

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

References

- [1] S.A. Wildman, G.M. Crippen, Prediction of physicochemical parameters by atomic contributions, *J. Chem. Inf. Comput. Sci.* 39 (1999) 868–873.
- [2] S.A. Wildman, G.M. Crippen, Evaluation of ligand overlap by atomic parameters, *J. Chem. Inf. Comput. Sci.* 41 (2001) 446–450.
- [3] S. Mankino, I.D. Kuntz, Automated flexible ligand docking and its application for database search, *J. Comp. Chem.* 18 (1997) 1812–1825.
- [4] G. Jones, P. Willett, R.C. Glen, A.R. Leach, R. Taylor, Development and validation of a genetic algorithm for flexible docking, *J. Mol. Biol.* 267 (1997) 727–748.
- [5] D.S. Goodsell, A.J. Olson, Automated docking of substrates to proteins by simulated annealing, *Proteins* 8 (1990) 195–202.
- [6] C. Wermuth, T. Langer, Pharmacophore identification, in: H. Kubinyi (Ed.), *3D QSAR in Drug Design: Theory, Methods and Applications*, ESCOM, Leiden, 1993, pp. 117–136.
- [7] V.E. Golender, E.R. Vorpapel, Computer-assisted pharmacophore identification, in: H. Kubinyi (Ed.), *3D QSAR in Drug Design: Theory, Methods and Applications*, ESCOM, Leiden, 1993, pp. 137–149.
- [8] G. Klebe, The use of composite crystal-field environments in molecular recognition and the de novo design of protein ligands, *J. Mol. Biol.* 237 (1994) 212–235.
- [9] G. Schneider, M.L. Lee, M. Stahl, P. Schneider, De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks, *J. Comput. Aid. Mol. Design* 14 (2000) 487–494.
- [10] J.H. Van Drie, Strategies for the determination of pharmacophoric 3D database queries, *J. Comput. Aid. Mol. Design* 11 (1997) 39–52.
- [11] W.P. Walters, M.T. Stahl, M.A. Murcko, Virtual screening—an overview, *Drug Disc. Today* 3 (1998) 160–178.
- [12] M.T. Stahl, M. Rarey, Detailed analysis of scoring functions for virtual screening, *J. Med. Chem.* 44 (2001) 1035–1042.
- [13] G.M. Crippen, VRI: 3D QSAR at variable resolution, *J. Comput. Chem.* 20 (1999) 1577–1585.
- [14] Molecular Operating Environment, 2000, Chemical Computing Group, Montreal, Canada.
- [15] J. Gasteiger, M. Marsili, Iterative partial equilization of orbital electronegativity—a rapid access to atomic charges, *Tetrahedron* 36 (1980) 3219–3228.
- [16] A.K. Ghose, G.M. Crippen, Geometrically feasible binding modes of a flexible ligand molecule at the receptor site, *J. Comput. Chem.* 6 (1985) 350–359.
- [17] S.A. Wildman, G.M. Crippen, 2002, manuscript in preparation.
- [18] S. Wold, A. Ruhe, H. Wold, W.J. Dunn, The colinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses, *SIAM J. Sci. Stat. Comput.* 5 (1984) 735–743.
- [19] R.D. Cramer III, D.E. Patterson, J.D. Bunce, Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins, *J. Am. Chem. Soc.* 110 (1988) 5959–5967.
- [20] E.A. Coats, The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods, *Perspect. Drug. Discov. Design* 12 (1998) 199–213.
- [21] M. Saunders, K.N. Houk, Y.D. Wu, W.C. Still, M. Lipton, G. Chang, W.C. Guida, Conformations of cycloheptadecane—a comparison of methods for conformational searching, *J. Am. Chem. Soc.* 112 (1990) 1419–1427.
- [22] H. Kubinyi, A general view on similarity and QSAR studies, in: H. van de Waterbeemd, B. Testa, G. Folkers (Eds.), *Proceedings of the 11th European Symposium on Quantitative Structure–Activity Relationships, Computer-Assisted Lead Finding and Optimization*, VCH, Basel, 1997, pp. 7–28.