Journal of Computer-Aided Molecular Design, 8 (1994) 323–340
ESCOM

J-CAMD 242

Extending the trend vector: The trend matrix and sample-based partial least squares

Robert P. Sheridan*, Robert B. Nachbar and Bruce L. Bush

Molecular Systems Department, Merck Research Laboratories, Rahway, NJ 07065, U.S.A.

Received 4 June 1993
Accepted 4 November 1993

Key words: Atom pairs; PLS; SAMPLS; Topological descriptors; QSAR

SUMMARY

Trend vector analysis [Carhart, R.E. et al., J. Chem. Inf. Comput. Sci., 25 (1985) 64], in combination with topological descriptors such as atom pairs, has proved useful in drug discovery for ranking large collections of chemical compounds in order of predicted biological activity. The compounds with the highest predicted activities, upon being tested, often show a several-fold increase in the fraction of active compounds relative to a randomly selected set. A trend vector is simply the one-dimensional array of correlations between the biological activity of interest and a set of properties or 'descriptors' of compounds in a training set. This paper examines two methods for generalizing the trend vector to improve the predicted rank order. The trend matrix method finds the correlations between the residuals and the *simultaneous* occurrence of descriptors, which are stored in a two-dimensional analog of the trend vector. The SAMPLS method derives a linear model by partial least squares (PLS), using the 'sample-based' formulation of PLS [Bush, B.L. and Nachbar, R.B., J. Comput.-Aided Mol. Design, 7 (1993) 587] for efficiency in treating the large number of descriptors. PLS accumulates a predictive model as a sum of linear components. Expressed as a vector of prediction coefficients on properties, the first PLS component is proportional to the trend vector. Subsequent components adjust the model toward full least squares. For both methods the residuals decrease, while the risk of overfitting the training set increases. We therefore also describe statistical checks to prevent overfitting. These methods are applied to two data sets, a small homologous series of disubstituted piperidines, tested on the dopamine receptor, and a large set of diverse chemical structures, some of which are active at the muscarinic receptor. Each data set is split into a training set and a test set, and the activities in the test set are predicted from a fit on the training set. Both the trend matrix and the SAMPLS approach improve the predictions over the simple trend vector. The SAMPLS approach is superior to the trend matrix in that it requires much less storage and CPU time. It also provides a useful set of axes for visualizing properties of the compounds. We describe a randomization method to determine the optimum number of PLS components that is very much faster for large training sets than leave-one-out cross-validation.

*To whom correspondence should be addressed.

INTRODUCTION

Although molecules are three-dimensional (3D) entities, relating their 3D properties to a biological activity is not always possible because of the difficulty of identifying the important conformations. Topological approaches, in which only the connection tables of molecules are considered, are therefore of great practical importance. Two independent aspects of topological methods need consideration: how to parse the connection tables of molecules into 'descriptors' and how to relate the descriptors to biological activity. Carhart et al. [1] and Nilakantan et al. [2] introduced two kinds of topological descriptors, the 'atom pair' and the 'topological torsion', respectively. These descriptors have two desirable features. First, they are easily computable from connection tables. Second, they are general enough that one can generate structure–activity relationships for sets of compounds with diverse chemical structures, but specific enough in the aggregate to discriminate among closely related isomers.

Carhart et al. [1] also described the 'trend vector' as a method of summarizing the correlation of discrete descriptors with biological activity and then using the summary to estimate the biological activity of new compounds. (The CASE system [3] has a similar philosophy.) The trend vector, used with the descriptors described above, has proved to be useful in a number of applications [4–7]. One application is the evaluation of chemical structures as candidates for synthesis. These structures can be sketched by the user [4] or stochastically assembled from fragments [5]. Another application is to use information from random screening to find more active compounds. Screening typically produces activity data on a few hundred chemically diverse compounds. Given a trend vector from these data, one can predict the activity of tens or hundreds of thousands of compounds in a corporate database. Upon testing the compounds with the highest predicted activities, one typically finds several-fold more actives per compound tested than in the random screening; often some of the actives will be from novel chemical classes [1,4,6].

The trend vector has a simple 1D mathematical representation. It is extremely robust, but fits the biological data rather coarsely. In this paper we describe two extensions to the mathematical form of the trend vector. In the first method, a 2D analog to the trend vector, a 'trend matrix', representing the correlation between descriptors, is added as a correction term. In the second, we apply SAMPLS, the newly described sample-based reformulation of partial least squares [8], to the realm of topological descriptors. We demonstrate that these extensions, in two tests, produce a better fit to the training set data without an unacceptable loss of robustness, thus yielding more accurate rankings of test sets.

METHODS

Review of the atom pair descriptor

In this paper, we will use only the atom pair descriptor, although the discussion could apply equally well to the topological torsion or any similar descriptor. An atom pair [1] is a three-part descriptor of the form:

atom type 1-(distance)-atom type 2

where 'distance' is the distance in bonds along the shortest path connecting the atoms. 'Atom type' includes information about element type, number of nonhydrogen neighbors, and number

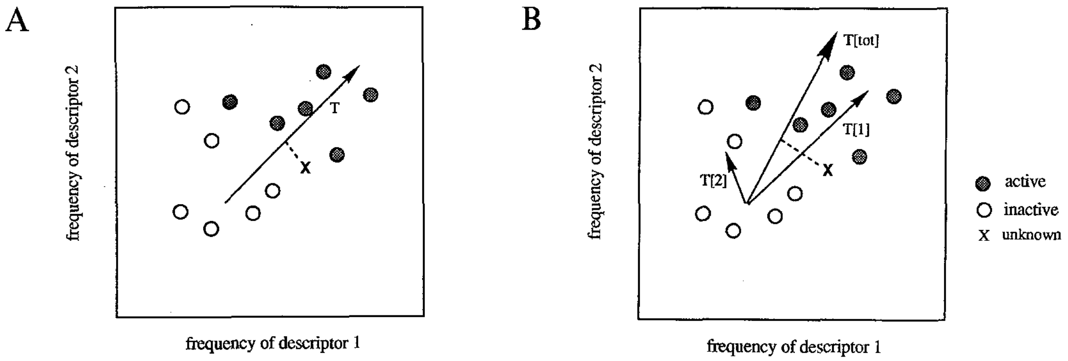


Fig. 1. (A) Schematic diagram of how a trend vector (T) is calculated from active and inactive compounds. The predicted activity of a new compound is its dot product on **T**. (B) Extension of trend vector calculation by sample-based partial least squares. **T**[1] and **T**[2] represent individual components in 'descriptor space'. **T**[tot] represents the sum of the individual components. The predicted activity of a new compound is its dot product on **T**[tot].

of π electrons. A chemical structure with n nonhydrogen atoms has n(n−1)/2 atom pairs, although many of these will be of the same type. It should be noted that atom pairs are meant to summarize the connection tables of sets of molecules; they are not explanatory in the sense that they are explicitly related to physical properties.

We follow the practice of precomputing the descriptors from a set of connection tables and storing them in a randomly accessible database [1]. The descriptors of relevant compounds are extracted from the database at run time.

Review of trend vector analysis

When a biological activity has been measured for a number of compounds, we can use the trend vector analysis to ask what structural features distinguish the more active compounds from the less active in the training set. Each compound can be represented as a point in a high-dimensional space where each dimension is associated with a distinct descriptor. Each compound is labeled with the biological activity of interest, in the simplest case classified as 'active' or 'inactive'. A two-descriptor case is illustrated in Fig. 1A.

A trend vector **T**, pointing from the inactive molecules toward the active ones, can be calculated by a formula analogous to that for calculating dipole moments, with 'activity' replacing 'charge':

$$T_k = \frac{1}{W_{tot}} \sum_{i}^{\text{compounds}} W_i A_i^{obs} X_{ik} \tag{1}$$

A_i^{obs} is the activity of compound i, normalized so that the weighted mean over the training set is 0.0 and the weighted standard deviation is 1.0. X_{ik} is a function of the number of times descriptor k (e.g., an atom pair of a particular type) occurs in compound i; in practice, X_{ik} is set to 1 if the descriptor k is present and to 0 if it is absent. We follow this treatment of **X** throughout. If there are many compounds from the same chemical class in a training set, weights W_i can be used to lower the contribution of individual compounds in that class, in effect retaining only one 'average' compound per class. W_{tot} is the sum of the weights.

Each coefficient T_k is proportional to the correlation coefficient of descriptor k with activity. Calculation of trend vectors is very rapid; training sets of hundreds or thousands of structures are easily handled. The length of **T**,

$$|\mathbf{T}| = \sqrt{\sum_{k}^{\text{descriptors}} T_k^2} \tag{2}$$

can be used to decide whether the structure–activity relationship described by the trend vector is statistically significant [1]. We produce randomized data sets by randomly reassigning the activities to the wrong compounds and renormalizing the activities. Presumably, there is no real structure–activity relationship in a randomized data set. The significance level of **T**, expressed in units of standard deviation above the mean, is calculated by comparing the |**T**| from the real data with the |**T**| values from 50 or more randomized data sets:

$$\text{significance level} = \frac{|\mathbf{T}_{\text{real}}| - \text{mean}\left(|\mathbf{T}_{\text{random}}|\right)}{\text{sd}\left(|\mathbf{T}_{\text{random}}|\right)} \tag{3}$$

This measures the confidence that the relationship between structures and activities in the training set is not due to chance. We usually take two or more standard deviations as an indication of statistical significance.

Given a significant trend vector, we predict the activity of some compound m not in the training set by finding the dot product of that compound onto **T**:

$$A_m^{\text{pred}} = \sum_{k}^{\text{descriptors}} T_k X_{mk} \tag{4}$$

where k runs over the distinct descriptors in the training set. Recognizing that the contribution of the absent descriptors to activity is unknown, we require that before a prediction can be made, a large fraction of the descriptors in compound i (typically 95% for atom pairs [1]) must be present in the training set from which the trend vector was derived. A_m^{pred} is in arbitrary units; for most trend vector applications, we are interested only in the *relative* activities of compounds. In this sense, the *direction* of the trend vector is important; not its magnitude. If necessary, A_m^{pred} can be rescaled to the original units by a weighted least-square fit of the A^{obs} vs. A^{pred} over the training set, yielding a scaling factor a.

Trend matrices

Trend matrices are rank-2 analogs of trend vectors that capture the correlation between descriptors. To calculate a trend matrix for a training set, we first generate the trend vector, calculate A^{pred} from Eq. 4, and find the residuals $R_i = A_i^{\text{obs}} - a\ A_i^{\text{pred}}$. The constant a rescales the arbitrary units of A^{pred} to the same units as the normalized A^{obs}. Vector **R** is then normalized. The trend matrix is:

$$M_{kk'} = \frac{1}{W_{\text{tot}}} \sum_{i}^{\text{compounds}} W_i R_i X_{ik} X_{ik'} \tag{5}$$

where k' > k. The indices k and k' represent separate descriptors. Since the values of **X** are either 0 or 1, including diagonal terms where k = k' is not necessary. There may be several million elements of **M**, and it is unwieldy to save them in a file. In practice, we save only the small fraction (say 1%) with the highest absolute values and use only these for the prediction.

The statistical significance of **M** can be monitored by evaluating the root sum of squares (rss) of its off-diagonal elements, which is the rank-2 analog of |**T**|. The rss from the real data is compared to the rss from randomized data sets by analogy with Eq. 3.

The predicted activity for compound m is:

$$A_m^{pred} = a \sum_{k}^{\text{descriptors}} T_k X_{mk} + b \sum_{k}^{\text{descriptors}} \sum_{k'>k}^{\text{descriptors}} M_{kk'} X_{mk} X_{mk'} \qquad (6)$$

A constant b is needed to rescale the trend matrix term back to the units of the residuals R_i.

Sample-based partial least squares

Partial least squares (PLS) is a statistical method described by Wold and others [9–11] that can be used even if the number K of properties (i.e. descriptors) exceeds the number N of samples (i.e. compounds). It reduces the explanatory data to a small number of components that correlate with some response (i.e. biological activity). This gives rise to a series of predictive models which fit the training set better and better, at the risk of overfitting. The first PLS component is proportional to the trend vector, so that PLS is a generalization of trend vector analysis.

The original PLS algorithm is formulated in terms of the N-by-K properties matrix **X**. A reformulation of Lindgren et al. [12] uses a K-by-K 'kernel' matrix, which substantially reduces storage and computing requirements when the number of compounds is very large ($N \gg K$). We will refer to both of these formulations of PLS as 'property-based'. Recently, we described the 'sample-based' reformulation of PLS called SAMPLS [8]. Unlike the property-based algorithms, SAMPLS is restricted to fitting a single property such as bioactivity, but it is much more computationally efficient than conventional PLS when the number of descriptors is very large ($K \gg N$). It is thus natural to apply SAMPLS to derive a PLS statistical model of biological activity for compounds represented by atom pairs; for a typical training set, the number of distinct atom pairs is usually much larger than the number of compounds.

SAMPLS begins by reducing **X** to an N-by-N matrix **C** which represents the covariances between compounds in the training set:

$$C_{ij} = \sum_{k}^{\text{descriptors}} X_{ik} X_{jk} \qquad (7)$$

In this context it should be noted that most other measures of molecular overlap are not suitable for a PLS analysis. One example is 'similarity' as defined by Carhart et al. [1]:

$$SIM_{ij} = \frac{2 \sum_{k}^{\text{descriptors}} \min(X_{ik}, X_{jk})}{\sum_{k}^{\text{descriptors}} X_{ik} + \sum_{k}^{\text{descriptors}} X_{jk}} \qquad (8)$$

(Unlike our definition everywhere else, X_{ik} in Eq. 8 represents the descriptor frequencies and not just the presence or absence of a descriptor.)

Once **C** has been computed, SAMPLS deals entirely in N-element vectors (see Ref. 8 for details). The description below is modified slightly from Ref. 8, to emphasize the individual components **v**[h] of the prediction coefficient vector **v**. The matrix algebra notation omits explicit W_i values; the reader should assume all inner products are weighted by W_i on each element. The routine SAMPLS_FIT can be described as follows:

```
Given: covariance matrix C; sample weights W; observed activities $A^{obs}$.
Calculate: for each PLS component h,
   residual activities (before the fit of component h) y[h];
   fitted activities (after the fit) $A^{fit}[h]$;
   prediction coefficient vector of component v[h].
Initialize: $y = A^{obs}$ (normalized and centered around 0), $A^{fit} = 0$
For components $h = 1,2,3,....,h_{max}$:
   Calculate fitting direction t and prediction coefficient vector v:
      Initialize this component:
         $s = C\ y$; center s;
         $u = y$
      If $(h > 1)$, project away preceding components:
         for $g = 1,2,..,(h-1)$
            $\alpha = (s^T t[g]) / (t[g]^T t[g])$
            $s = s - \alpha\ t[g]$
            $u = u - \alpha\ v[g]$
      Endif
      Set $t = s$
   Scale t for best least-squares fit to y:
      $\beta = (t^T\ y) / (t^T t)$
   Save for subsequent fit and for prediction:
      $t[h] = t$
      $v[h] = \beta\ u$
      $y[h] = y$
   Update residual y; update and save fitted activity $A^{fit}$:
      $y = y - \beta\ t$
      $A^{fit} = A^{fit} + \beta\ t$
      $A^{fit}[h] = A^{fit}$
   (No need to update X or C)
End
```

In our implementation, generating a PLS model is done in two phases. In the validation phase, one determines how many PLS components are statistically significant. In practice, since only the first few components are likely to be significant, the user chooses $h_{max} \ll N$. The data are varied in some fashion as discussed below, and the statistical behavior of each added component is analyzed. In the second model generation phase, the user then chooses h_{incl}, the number of components to be included in the final model, and reruns SAMPLS_FIT with $h_{max} = h_{incl}$.

To predict the activity of compounds not in the training set, the PLS model can be expressed compactly as an N-element vector **v**[tot] of prediction coefficients in 'compound space', or equivalently as a single K-element vector **T**[tot] of prediction coefficients in 'descriptor space'. Each component **T**[h] can be calculated from the corresponding **v**[h] as a linear combination of rows of **X** over the compounds in the training set:

$$\mathrm{T[h]_k} = \frac{1}{\mathrm{W_{tot}}} \sum_{\mathrm{i}}^{\text{compounds}} \mathrm{W_i\, v[h]_i\, X_{ik}} \tag{9}$$

v[1] is proportional to $\mathbf{A}^{\mathrm{obs}}$, so **T**[1] is a scalar multiple of the simple trend vector. Subsequent components span additional dimensions in descriptor space (although they are not orthogonal in this space). The **T**[h] values are summed to obtain the final prediction coefficients **T**[tot] of the PLS model:

$$\mathbf{T}[\mathrm{tot}] = \mathbf{T}[1]+\mathbf{T}[2]+\mathbf{T}[3]+\ldots+\mathbf{T}[\mathrm{h_{incl}}] \tag{10}$$

The PLS model calculated in this way is identical to that from property-based formulations of PLS. The predicted activity of an unknown compound m is its dot product on **T**[tot]:

$$\mathrm{A_m^{pred}} = \sum_{\mathrm{k}}^{\text{descriptors}} \mathrm{T[tot]_k\, X_{mk}} \tag{11}$$

This is shown schematically in Fig. 1B for a two-component problem. If $\mathrm{h_{incl}}=1$, Eq. 11 reduces to Eq. 4. We have confirmed that when Eq. 11 is applied to the training set, the predicted activities for a given $\mathrm{h_{incl}}$ are proportional to the $\mathbf{A}^{\mathrm{fit}}$ values from SAMPLS_FIT.

We have employed two methods to determine the appropriate number of components $\mathrm{h_{incl}}$. The conventional method is cross-validation. The simplest way to perform cross-validation is the leave-one-out procedure, where each compound in turn is eliminated from the training set, and its activity is predicted by a fit to all the other compounds. In our validation phase, we leave out compound i by setting $\mathrm{W_i}$ to zero for this compound before rerunning SAMPLS_FIT; the prediction for i after h components is the $\mathrm{A^{fit}[h]_i}$ from SAMPLS-FIT. The cross-validated $\mathrm{r^2}$ summarizes the agreement between observed and predicted activity:

$$\text{cross-validated } \mathrm{r^2} \text{ after h components} = 1.0 - \frac{\sum_{\mathrm{i}}^{\text{compounds}} \mathrm{W_i(A^{fit}[h]_i - A_i^{obs})^2}}{\sum_{\mathrm{i}}^{\text{compounds}} \mathrm{W_i(A_i^{obs} - A_{mean})^2}} \tag{12}$$

where $\mathrm{A_{mean}}$ is the weighted mean of $\mathbf{A}^{\mathrm{obs}}$. The optimum $\mathrm{h_{incl}}$ is usually taken as the value of h where the cross-validated $\mathrm{r^2}$ is maximal.

Although SAMPLS provides an extremely efficient method for cross-validation, independent of the number of descriptors [8], this procedure can be time-consuming for large training sets. A randomization approach, analogous to that used for trend vectors, provides an alternative way to determine how many PLS components to keep. This method needs only a single run of SAMPLS_FIT in the validation phase. For each component h we normalize **y**[h], the residuals *before* the fit of component h, and calculate vectors **S**[h]:

$$\mathrm{S[h]_k} = \frac{1}{\mathrm{W_{tot}}} \sum_{\mathrm{i}}^{\text{compounds}} \mathrm{W_i\, y[h]_i\, X_{ik}} \tag{13}$$

$\mathrm{S[h]_k}$ is the correlation over the training set of descriptor k with **y**[h]. The S's, unlike the T's, form an orthogonal set. (They are proportional to the 'X-weights' of property-based formulations; see

Refs. 9–12.) Formally, for h > 1 the matrix **X** in Eq. 13 should be the updated matrix after components 1,2,..,h – 1. However, the construction of the residuals **y**[h] ensures that Eq. 13 yields the same results, independent of whether **X** is updated (see Ref. 8).

We produce randomized residuals by randomly reassigning the residuals to the wrong compounds and renormalizing **y**[h]. The significance level of **S**[h] is calculated, by analogy with Eq. 3, by comparing |**S**[h]| from the original residuals vs. that from 50 or more sets of randomized residuals. It measures the confidence that the relationship between structures and residuals in component h is not due to chance. A component for which there is no such real relationship is probably fitting noise. Since **y**[1] is the normalized $\mathbf{A}^{obs}$, the significance level of **S**[1] is identical to that for a simple trend vector. It should be stressed that we are determining the significance of *each* component in turn, while the cross-validated r^2 is used to determine the optimum *total* number of components. Aside from speed, the randomization method has an advantage over cross-validation in that it measures statistical significance directly, whereas cross-validation measures the predictive ability (although the significance may be estimated from the results of another study [13]).

RESULTS

We have chosen two data sets as examples. For both, we split the data sets into a training and a test set. Using each prediction method, we compare the measures of reliability within the training set (cross-validation or randomization) with the success of the models in ranking compounds within the test set.

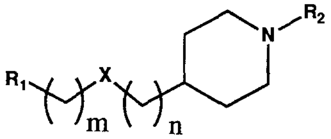
1. Dopaminergic activity of disubstituted piperidines

Our first example is small enough that both trend matrices and SAMPLS can be tried, and homogeneous enough that the descriptors can be interpreted easily. A set of disubstituted piperidines were synthesized by Gilligan et al. [14] as sigma receptor ligands. They tested these compounds for four types of biological activity. The affinity of these compounds for the dopamine D2 receptor has the largest range (~3 orders of magnitude) of all the activities. We use $-\log(IC_{50})$ for this receptor as our observed activity. We alternately assigned the compounds, taken in the order they are listed in Ref. 14, to a training set and a test set of equal size; see Table 1. All compounds have a weight of 1.0. The training set and test set have 58 compounds each, with mean activities of 5.61 ± 0.73 and 5.68 ± 0.73, respectively. The number of distinct atom pairs in the training set is 724.

Trend vector

The total CPU time for the trend vector calculation on the training set, for 50 randomization trials, was 10 s on a VAX 8840. (This does not include the time required to extract the descriptors from the database.) The length of the trend vector was 1.58 vs. 0.89 ± 0.13 for the randomized sets, giving a good significance level of 5.2 sd. The most important atom pairs of the trend vector are shown in Fig. 2B. Clearly, the substituent R_1 = 4-F-Ph is most statistically associated with activity (**T** descriptors 1–10) and there is a preference for m = 0 and n = 1 as the number of $-CH_2$-spacers (hence descriptors 4–10). Substituent $R_2 = CH_2$-cPr is associated with inactivity (descriptor 724).

TABLE 1
BINDING OF DISUBSTITUTED PIPERIDINES (FROM Ref. 10) TO DOPAMINE RECEPTORS



No.	X	m	n	R$_1$	R$_2$	$-\log(IC_{50})$
Training set						
1	CO	0	1	4-F-Ph	CH$_2$Ph	5.8
6b	CO	0	1	4-MeO-Ph	CH$_2$Ph	5.2
6d	CO	0	1	4-HO-Ph	CH$_2$Ph	5.9
6f	CO	0	1	4-HOCH$_2$-Ph	CH$_2$Ph	5.0
6h	CO	0	1	4-MeS(O)-Ph	CH$_2$Ph	5.0
6j	CO	0	1	4-Cl-Ph	CH$_2$-cPr	5.7
6l	CO	0	1	4-*t*Bu-Ph	CH$_2$-cPr	5.8
6n	CO	0	1	4-CF$_3$-Ph	CH$_2$-cPr	5.0
6p	CO	0	1	4-NH$_2$-Ph	CH$_2$-cPr	5.0
6r	CO	0	1	4-F-Ph	CH$_2$Ph-*p*-CF$_3$	6.7
6t	CO	0	1	4-F-Ph	(CH$_2$)$_2$-3-indolyl	8.2
6v	CO	0	1	4-F-Ph	(CH$_2$)$_2$Ph	7.4
6x	CO	0	1	4-F-Ph	(CH$_2$)$_2$Ph-*p*-CF$_3$	8.0
7a	CHOH	0	1	4-F-Ph	(CH$_2$)$_2$-cPr	5.0
7c	CHOH	0	1	4-MeS-Ph	CH$_2$Ph	5.1
7e	CHOH	0	1	4-CF$_3$-Ph	CH$_2$Ph	5.0
7g	CHOH	0	1	2-naphthyl	CH$_2$Ph	5.5
7i	CHOH	0	1	2-furyl	CH$_2$Ph	5.0
10a	CHOH	1	0	4-F-Ph	CH$_2$Ph	6.1
10c	CHOH	1	0	4-F-Ph	(CH$_2$)$_3$Ph	5.8
11a	CO	1	0	4-F-Ph	CH$_2$Ph	5.3
11c	CO	1	1	4-F-Ph	CH$_2$Ph	6.0
18a	O	0	1	4-F-Ph	CH$_2$-cPr	6.4
18c	O	0	1	4-MeO-Ph	CH$_2$-cPr	5.0
18e	O	0	1	4-HOCH$_2$Ph	CH$_2$-cPr	5.0
18g	O	0	1	4-MeCH(OH)-Ph	CH$_2$-cPr	5.0
18i	O	0	1	F$_5$Ph	CH$_2$-cPr	5.2
18k	O	0	1	4-MeS-Ph	CH$_2$-cPr	5.7
18m	O	0	1	4-NO$_2$-Ph	CH$_2$-cPr	5.0
18o	O	0	1	4-MeCOPh	CH$_2$-cPr	6.3
18r	O	0	1	4-(4'-F-Ph)-Ph	CH$_2$-cPr	5.0
18t	O	0	1	Ph	CH$_2$-cPr	5.5
18v	O	0	1	3,4-Cl$_2$-Ph	CH$_2$-cPr	5.8
18x	O	0	1	4-EtN-Ph	CH$_2$-cPr	5.0
18z	O	0	1	4-F-Ph	CH$_2$-MeCl$_2$-cPr	6.7
18ab	O	0	1	4-Cl-Ph	CH$_2$Ph	5.4
18ad	O	0	1	4-MeO-Ph	CH$_2$Ph	5.0
18af	O	0	1	4-F-Ph	CH$_2$-Ph-*p*-F	6.0
18ah	O	0	1	4-F-Ph	CH$_2$-2-naphthyl	5.4
18aj	O	0	1	4-F-Ph	(CH$_2$)$_2$-Ph-*p*-Cl	6.4
18al	O	1	1	4-F-Ph	CH$_2$Ph	5.4
18an	O	1	1	4-Ph-Ph	CH$_2$Ph	5.7
18ap	O	1	1	4-MeCO$_2$-Ph	CH$_2$Ph	6.0
18ar	O	1	1	4-F-Ph	(CH$_2$)$_3$Ph	5.0
18at	O	1	1	4-F-Ph	CH$_2$Ph-*p*-Cl	6.3

TABLE 1 (continued)

No.	X	m	n	R_1	R_2	$-\log(IC_{50})$
18av	O	1	1	4-F-Ph	CH_2Ph-*p*-OH	5.5
18ax	O	1	1	4-F-Ph	CH_2-4-pyridyl	5.0
18az	O	1	1	4-F-Ph	CH_2-2-naphthyl	6.5
18bb	O	1	1	4-F-Ph	$(CH_2)_4CH_3$	5.9
18bd	O	3	0	Ph	CH_2Ph	5.1
18bf	O	4	1	Ph	CH_2Ph	5.4
18bh	O	1	2	4-*t*Bu-Ph	CH_2Ph	5.2
18bj	O	0	1	4-pyridyl	CH_2-cPr	5.0
18bl	O	0	1	2-pyrimidyl	CH_2-cPr	5.4
18bn	O	0	1	5-indolyl	CH_2-cPr	5.6
18bp	O	0	1	cyclohexyl	CH_2Ph	5.0
18br	O	0	1	3-pyridyl	CH_2Ph	5.0
22	SO	0	1	4-F-Ph	CH_2-cPr	5.0
Test set						
6a	CO	0	1	4-CF_3-Ph	CH_2Ph	5.3
6c	CO	0	1	4-MeS-Ph	CH_2Ph	6.1
6e	CO	0	1	4-Ph-Ph	CH_2Ph	5.5
6g*	CO	0	1	4-$MeSO_2$-Ph	CH_2Ph	5.0
6i	CO	0	1	4-F-Ph	CH_2-cPr	5.8
6k	CO	0	1	4-MeO-Ph	CH_2-cPr	5.7
6m	CO	0	1	4-Ph-Ph	CH_2-cPr	5.8
6o	CO	0	1	4-NMe_2-Ph	CH_2-cPr	5.0
6q*	CO	0	1	4-CN-Ph	CH_2-cPr	5.0
6s	CO	0	1	4-F-Ph	CH_2Ph-*p*-F	6.2
6u	CO	0	1	4-F-Ph	$(CH_2)_2Ph$-*p*-F	7.2
6w	CO	0	1	4-F-Ph	$(CH_2)_2Ph$-*p*-Cl	7.9
6y	CO	0	1	4-F-Ph	$(CH_2)_2$-cPr	6.7
7b	CHOH	0	1	4-F-Ph	CH_2Ph	5.2
7d	CHOH	0	1	4-MeO-Ph	CH_2Ph	5.0
7f	CHOH	0	1	4-F-Ph	$(CH_2)_2Ph$	6.1
7h*	CHOH	0	1	2-thienyl	CH_2Ph	5.4
10b	CHOH	1	0	Ph	CH_2Ph	5.1
10d	CHOH	1	0	4-F-Ph	$(CH_2)_4Ph$	6.0
11b*	CO	1	0	4-F-Ph	CH_2-4-pyridyl	5.0
11d	CO	1	0	4-F-Ph	$(CH_2)_3COPh$-*p*-F	7.3
18b	O	0	1	4-Cl-Ph	CH_2-cPr	5.0
18d	O	0	1	4-Ph-Ph	CH_2-cPr	5.0
18f	O	0	1	4-*t*Bu-Ph	CH_2-cPr	5.5
18h	O	0	1	3,4-F_2-Ph	CH_2-cPr	5.5
18j*	O	0	1	3,4,5-$(MeO)_3$-Ph	CH_2-cPr	5.0
18l*	O	0	1	4-$MeSO_2$-Ph	CH_2-cPr	5.0
18n*	O	0	1	4-CN-Ph	CH_2-cPr	5.0
18q*	O	0	1	4-PhO-Ph	CH_2-cPr	5.0
18s*	O	0	1	4-(4'-MeO-Ph)-Ph	CH_2-cPr	5.1
18u*	O	0	1	3-Me_2N-Ph	CH_2-cPr	5.1
18w*	O	0	1	2,4-Cl_2-Ph	CH_2-cPr	7.0
18y*	O	0	1	4-F-Ph	CH_2-(2'-Me)-cPr	6.0
18aa	O	0	1	4-F-Ph	CH_2Ph	6.1
18ac	O	0	1	4-NO_2-Ph	CH_2Ph	5.4
18ae	O	0	1	4-CF_3-Ph	CH_2Ph	5.7
18ag*	O	0	1	4-F-Ph	$(CH_2)_2Ph$-*p*-OMe	5.9
18ai	O	0	1	4-F-Ph	CH_2-4-pyridyl	5.6

TABLE 1 (continued)

No.	X	m	n	R_1	R_2	$-\log(IC_{50})$
18ak	O	0	1	4-F-Ph	$(CH_2)_2$-cPr	6.4
18am*	O	1	1	4-MeO-Ph	CH_2Ph	4.6
18ao	O	1	1	Ph	CH_2Ph	5.0
18aq	O	1	1	4-F-Ph	$(CH_2)_2Ph$	6.3
18as*	O	1	1	4-F-Ph	CH_2Ph-*p*-CO_2Me	5.7
18au	O	1	1	4-F-Ph	CH_2Ph-*p*-Ph	6.1
18aw*	O	1	1	4-F-Ph	CH_2Ph-*p*-OBz	6.1
18ay	O	1	1	4-F-Ph	CH_2-cyclohexyl	5.5
18ba	O	1	1	4-F-Ph	CH_2-1-naphthyl	6.7
18bc	O	0	2	4-F-Ph	CH_2Ph	6.9
18be	O	3	1	Ph	CH_2Ph	6.1
18bg	O	5	1	Ph	CH_2Ph	5.7
18bi	O	0	1	2-naphthyl	CH_2-cPr	5.5
18bk	O	0	1	4-quinolinyl	CH_2-cPr	5.3
18bm	O	0	1	2-pyridyl	CH_2-cPr	5.3
18bo	O	0	1	2-naphthyl	CH_2Ph	6.0
18bq	O	0	1	2-quinolinyl	CH_2Ph	5.2
18bs	O	0	1	cPr	CH_2Ph	5.0
18bt*	S	0	1	4-F-Ph	$(CH_2)_2$-cPr	5.8
23*	SO_2	0	1	4-F-Ph	$(CH_2)_2$-cPr	5.0

* These compounds could not be predicted according to the requirement that at least 95% of the distinct atom pairs in the compound to be predicted must be present in the training set.

Trend vector plus trend matrix

This run took 424 s for 50 randomization trials. The results for the initial vector calculation were identical to those above. The rss of **M** for the training set was 10.62 vs. 6.74 ± 0.71 for the randomized sets, for a significance level of 5.4 sd. Thus, we find a significant correlation of cross terms in **M** with the residuals left after fitting by **T**. The most positive cross terms indicate that compounds with X = O and $R_1 = CH_2$-cPr are more active than would be predicted from **T** alone (**M** descriptors 1–10). One percent (2617) of the most significant elements of **M** were written to a file to be used for prediction.

SAMPLS

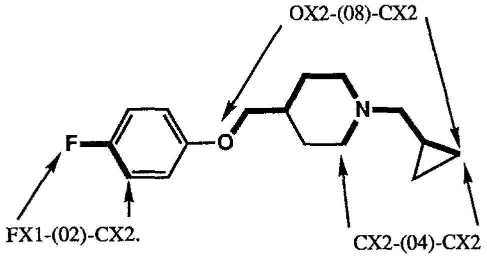
For this example, a run using the leave-one-out method for validation with $h_{max} = 10$ took 18 s. A run using randomization for validation took 233 s for 50 trials. In both runs the calculation of the covariances took 2 s of the total time. The cross-validated r^2 and significance levels are:

Total components	1	2	3	4	5	6	7	8	...
Cross-validated r^2	0.308	0.409	0.426	0.419	0.403	0.387	0.373	0.314	...
Significance level (sd)	5.2	2.5	−0.7	−1.3	−0.5	−2.7	−2.3	−2.2	...

The cross-validated r^2 indicates that using three total components is optimum. The significance levels show that the relationship between residuals and structures in components 1 and 2 is probably not due to chance, but thereafter the components are fitting noise. For completeness, we will use $h_{incl} = 3$ in the following discussion.

The vectors **T**[1], **T**[2] and **T**[3] are listed in Fig. 2C. **T**[1] is proportional to the trend vector derived before. **T**[tot] has a number of important long-range descriptors, indicating that com-

A



B

	T		M
1	0.242 F X1 –(02)–C X2.	**1**	0.207 O X2 –(01)–C X3. C X2 –(04)–C X2
2	0.234 F X1 –(01)–C X3.	**2**	0.207 O X2 –(08)–C X2 C X2. –(10)–C X2
3	0.234 F X1 –(04)–C X3.	**3**	0.207 O X2 –(08)–C X2 C X2. –(11)–C X2
4	0.231 F X1 –(07)–C X3	**4**	0.206 C X2 –(04)–C X2 C X3. –(07)–C X2
5	0.231 F X1 –(08)–C X2	**5**	0.201 O X2 –(01)–C X3. C X2 –(07)–C X2
6	0.231 F X1 –(11)–C X2	**6**	0.199 C X2 –(04)–C X2 O X2 –(08)–C X2
7	0.231 N X3 –(10)–F X1	**7**	0.199 C X2 –(07)–C X2 O X2 –(08)–C X2
8	0.229 F X1 –(14)–C X2.	**8**	0.199 O X2 –(01)–C X2 O X2 –(08)–C X2
9	0.224 F X1 –(15)–C X2.	**9**	0.199 O X2 –(08)–C X2 C X3. –(09)–C X2
10	0.221 F X1 –(06)–C X2	**10**	0.199 O X2 –(04)–C X2 O X2 –(08)–C X2
	o		o
	o		o
715	–0.100 O X2 –(07)–C X3	**2608**	–0.210 N X3 –(02)–C X3. C X2. –(13)–C X2.
716	–0.111 O X1 –(03)–C X2.	**2609**	–0.210 N X3 –(03)–C X2. C X2. –(13)–C X2.
717	–0.112 N X3 –(02)–C X3	**2610**	–0.210 C X3 –(06)–C X2. C X3. –(11)–C X2.
718	–0.112 C X3 –(05)–C X3	**2611**	–0.210 N X3 –(03)–C X2. C X3. –(11)–C X2.
719	–0.114 O X1 –(07)–C X2	**2612**	–0.214 C X3. –(05)–C X3 C X3. –(11)–C X2.
720	–0.115 O X1 –(02)–C X3.	**2613**	–0.214 C X3. –(05)–C X3 C X2. –(13)–C X2.
721	–0.119 C X2 –(08)–C X2	**2614**	–0.214 C X3 –(06)–C X2. C X3. –(14)–C X2.
722	–0.127 C X3 –(10)–C X2.	**2615**	–0.215 C X3 –(06)–C X2. C X3. –(13)–C X2.
723	–0.157 C X3 –(03)–C X2	**2616**	–0.215 C X3. –(03)–C X2 C X3. –(05)–C X3
724	–0.167 C X3 –(06)–C X2	**2617**	–0.230 C X3 –(06)–C X2. C X2. –(13)–C X2.

C

PLS	**T**[1]	**T**[2]	**T**[3]	**T**[tot]
1	0.153 F X1 –(02)–C X2.	0.104 C X2 –(04)–C X2	0.143 C X3. –(12)–C X3.	0.378 C X3. –(12)–C X3.
2	0.147 F X1 –(01)–C X3.	0.103 C X3. –(12)–C X3.	0.127 C X3. –(09)–C X3.	0.365 C X3. –(09)–C X3.
3	0.147 F X1 –(04)–C X3.	0.101 C X3. –(09)–C X3.	0.127 C X2 –(07)–C X1	0.325 N X3 –(03)–C X3.
4	0.146 F X1 –(07)–C X3	0.098 N X3 –(03)–C X3.	0.121 O X1. –(09)–C X3.	0.310 O X1. –(09)–C X3.
5	0.146 F X1 –(08)–C X2	0.095 C X2. –(10)–C X2	0.116 C X3 –(07)–C X2.	0.302 F X1 –(13)–C X3.
6	0.146 F X1 –(11)–C X2	0.095 C X3. –(08)–C X3	0.115 N X3 –(03)–C X3.	0.259 O X1. –(08)–C X2
7	0.146 N X3 –(10)–F X1	0.092 F X1 –(06)–C X2	0.103 N X3 –(04)–C X2.	0.254 C X3. –(11)–C X2
8	0.144 F X1 –(14)–C X2.	0.092 C X3. –(11)–C X2	0.098 C X2 –(09)–C X1	0.251 F X1 –(16)–C X3.
9	0.141 F X1 –(15)–C X2.	0.089 O X1. –(09)–C X3.	0.095 C X3 –(09)–C X2.	0.247 F X1 –(06)–C X2
10	0.140 F X1 –(06)–C X2	0.082 C X3. –(09)–C X2	0.094 F X1 –(16)–C X3.	0.226 O X1. –(06)–F X1
	o	o	o	o
	o	o	o	o
715	–0.063 O X2 –(07)–C X3	–0.071 O X1 –(07)–C X2	–0.089 O X1. –(02)–C X2	–0.152 C X3 –(06)–C X2.
716	–0.070 O X1 –(03)–C X2.	–0.073 O X1. –(09)–C X2.	–0.089 O X1. –(04)–C X2	–0.152 N X3 –(11)–F X1
717	–0.071 N X3 –(02)–C X3	–0.073 C X3. –(08)–C X2.	–0.097 N X3 –(11)–F X1	–0.164 C X3 –(07)–C X2
718	–0.071 C X3 –(05)–C X3	–0.075 C X3. –(03)–C X2	–0.097 F X1 –(08)–C X3	–0.175 C X3. –(05)–C X3
719	–0.072 O X1 –(07)–C X2	–0.083 C X2 –(08)–C X2	–0.099 F X1 –(09)–C X2	–0.181 N X3 –(02)–C X3.
720	–0.073 O X1 –(02)–C X3.	–0.112 C X3 –(06)–C X2.	–0.104 F X1 –(03)–C X2.	–0.189 C X2. –(10)–C X2.
721	–0.075 C X2 –(08)–C X2	–0.116 C X2. –(10)–C X2.	–0.105 F X1 –(10)–C X2	–0.204 F X1 –(13)–C X3
722	–0.080 C X3 –(10)–C X2.	–0.117 C X3. –(05)–C X3	–0.127 F X1 –(13)–C X3	–0.217 N X3 –(03)–C X2.
723	–0.099 C X3 –(03)–C X2	–0.131 N X3 –(02)–C X3.	–0.131 F X1 –(12)–C X2	–0.222 F X1 –(14)–C X2
724	–0.106 C X3 –(06)–C X2	–0.137 N X3 –(03)–C X2.	–0.137 F X1 –(14)–C X2	–0.244 C X2 –(08)–C X2

Fig. 2. List of atom pairs in order of coefficient value for disubstituted piperidines. Atom nomenclature includes element type, number of nonhydrogen neighbors (X1,X2,X3, etc.) and number of π electrons (' ' = 0, '.' = 1, ':' = 2, etc.). The number of bonds separating the atoms is given in parentheses. (A) Structure of compound with sample atom pairs labeled. (B) Trend vector plus trend matrix corrections. (C) Trend vector components derived from sample-based PLS. For this figure, T[1] was scaled to have a length of 1.0; other T's were scaled relative to T[1].

pounds with $R_1 = 4$-F-Ph, $m = 0$, and $R_2 = -(CH_2)_2$-aryl are most likely to be active. Projection of the training set into the 3D space defined by the first three components is shown in Fig. 3. It is easily seen that the projection along **T**[tot] better separates active and inactive compounds than does the projection along **T**[1], which corresponds to the simple trend vector. Another way of saying this is that the 'discriminant plane' through the data is nearly perpendicular to **T**[tot].

Predictions

Predictions were made of the training and test sets using each method, i.e., by Eqs. 4, 6 and 11. (Prediction of the training set is, of course, not a true prediction but a reflection of the previous fit.) Given the usual requirement that 95% of the descriptors in a compound to be predicted be present in the training set [1], only 40 of the 58 compounds in the test set can be predicted. The results of the predictions are summarized in Table 2 and shown graphically in Fig. 4. In Table 2, we show the results for $h_{incl} = 1$, 2 and 3. As expected, application of the trend matrix and PLS results in a large improvement over **T** alone in the fit of the training set and a smaller, but significant, improvement in the prediction of the test set.

2. 'Muscarinic activity' in a large database

Our second example is an attempt to simulate, using published data, the common situation where the investigator has assay results on a large set of diverse chemical structures. Often the activity is reported only qualitatively, i.e., 'active' or 'inactive'. One can calculate a trend vector from these data and then use the trend vector to select compounds out of a large database for further testing. We used compounds from the MACCS Drug Data Report (MDDR) [15], which is a licensed database of about 29000 drug-like structures, compiled from the open literature. Most structures have associated text fields, including 'activity category' and '(mode of) action'. We found in the MDDR database 381 compounds that refer to the muscarinic receptor in the 'action' data field. These we assigned an activity of 1.0. (We did not discriminate among receptor subtypes or between agonists and antagonists.) We randomly selected about 2000 MDDR com-

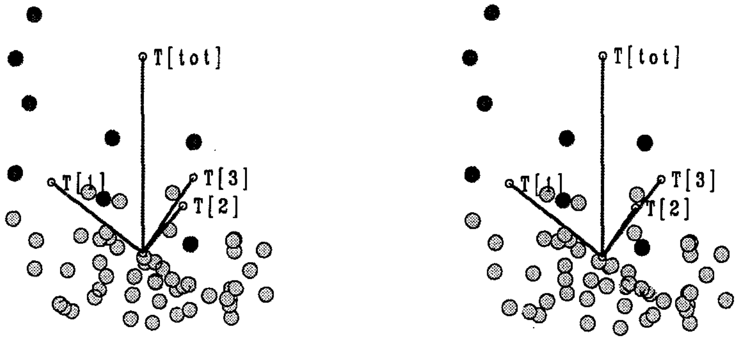


Fig. 3. The projection of each compound in the training set of disubstituted piperidines in the coordinate space of the three components T[1], T[2], T[3]. (The T[1] coordinate of compound i is the dot product of the descriptors of i on the unit vector pointing in the direction T[1], etc.) The sum of the vectors T[1]–T[3] is along the direction T[tot]. T[1] and T[tot] are in the plane of the page. The lengths of the axes reflect the relative lengths of the T's. Although this plot is on orthogonal axes, the T's are not really orthogonal in descriptor space. For purposes of this figure, compounds are arbitrarily divided into 'actives' (dark circles) and 'inactives' (light circles). The cutoff is at $-\log(IC_{50}) = 6.3$, which is one sd above the mean activity.

TABLE 2

CORRELATION BETWEEN PREDICTED AND OBSERVED ACTIVITIES FOR DISUBSTITUTED PIPERIDINES IN TABLE 1

Method[a]	TV = ^{1}PLS	^{2}PLS	^{3}PLS	TV + TM
Training set	0.73	0.85	0.93	0.83
Test set	0.77	0.81	0.81	0.81

[a] TV = trend vector method; TV+TM = trend vector plus trend matrix method. PLS = SAMPLS fit, followed by transformation of the results to descriptor space. ^{1}PLS = fit with a total of one component, ^{2}PLS with two components, etc.

pounds for which the muscarinic receptor is *not* mentioned in the 'action' field and assigned these an activity of 0.0. We split this large set by randomly selecting compounds, with a probability of 0.3, to be in the training set; the remainder were added to the test set. To simulate diverse sets, we eliminated compounds within each set so that no two compounds had a similarity greater than 0.6 (Eq. 8). The final training set contained 502 compounds, 55 of which were actives. The final test set comprised 919 compounds, 69 of which were actives. The training set contained 4525 distinct descriptors. Again, all the compounds were given a weight of 1.0.

Trend vector

This calculation took 31 s for 50 randomization trials. The length of the trend vector was 1.29, vs. 0.52 ± 0.04 for the randomized sets, giving a very high significance level of 19.2. The most significant atom pairs of the trend vector are shown in Fig. 5 as **T**[1]. This set is not straightforward to interpret, because it is an average of many classes of compounds. However, it is clear that aliphatic tertiary amines are important for activity.

Trend vector plus trend matrix

Trend matrix calculations are not feasible for a problem of this size. Given 4525 atom pair descriptors, it would require storing a matrix with 10.2 million elements.

SAMPLS

This run took 8104 s for a leave-one-out validation with $h_{max} = 10$. A run using randomization

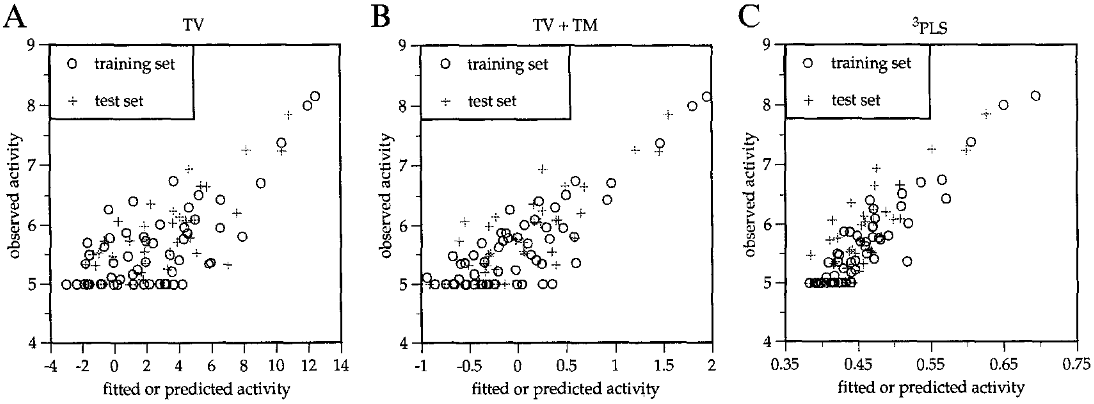
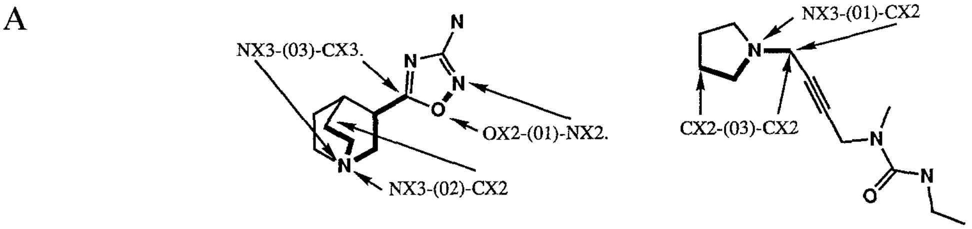


Fig. 4. Observed vs. predicted activities for three methods applied to the disubstituted piperidines. The units for the predicted activity are arbitrary. (A) Trend vector method; (B) trend vector plus trend matrix; (C) a fit from three trend vector components.



B

PLS	T[1]	T[2]	T[3]
1	0.110 N X3 –(02)–C X2	0.416 N X3 –(02)–C X2	0.596 N X3 –(02)–C X2
2	0.107 N X3 –(01)–C X2	0.389 N X3 –(01)–C X2	0.578 N X3 –(01)–C X2
3	0.079 C X2 –(03)–C X2	0.355 C X2 –(02)–C X2	0.470 N X2 –(03)–C X3.
4	0.078 C X2 –(02)–C X2	0.320 C X2 –(03)–C X2	0.458 N X3 –(01)–C X1
5	0.064 N X3 –(01)–C X1	0.267 N X3 –(01)–C X1	0.390 O X2 –(01)–N X2.
6	0.062 N X2. –(04)–C X2	0.220 C X2 –(01)–C X2	0.365 C X2 –(02)–C X2
7	0.056 N X2. –(05)–C X2	0.214 O X2 –(01)–N X2.	0.333 C X3. –(05)–C X2.
8	0.051 C X2 –(01)–C X2	0.208 O X2 –(04)–C X2	0.328 N X3 –(05)–C X3.
9	0.050 O X2 –(01)–N X2.	0.206 C X2 –(05)–C X1	0.318 O X1. –(06)–C X2.
10	0.044 N X3 –(03)–C X3.	0.190 O X2 –(03)–C X2	0.297 C X2 –(03)–C X2
	○	○	○
	○	○	○
4516	–0.099 C X3. –(08)–C X3.	–0.197 C X3. –(07)–C X2.	–0.240 C X3. –(01)–C X3.
4517	–0.099 O X1. –(07)–C X3.	–0.203 C X2. –(02)–C X2.	–0.243 C X3 –(08)–C X2
4518	–0.099 O X1. –(04)–C X3.	–0.203 C X3. –(05)–C X3.	–0.243 N X3 –(02)–N X2.
4519	–0.101 C X3. –(08)–C X2.	–0.203 N X3 –(01)–C X3.	–0.256 O X1. –(03)–C X2
4520	–0.102 C X3. –(06)–C X3.	–0.216 O X1. –(04)–C X3.	–0.260 C X2 –(04)–C X2
4521	–0.103 C X3. –(03)–C X2.	–0.222 C X3. –(06)–C X2.	–0.274 N X3 –(01)–C X3.
4522	–0.104 C X3. –(05)–C X3.	–0.227 C X2. –(01)–C X2.	–0.282 C X3. –(03)–C X2.
4523	–0.106 C X3. –(09)–C X2.	–0.248 C X3. –(03)–C X3.	–0.307 C X3 –(07)–C X2
4524	–0.108 C X3. –(06)–C X2.	–0.290 C X3. –(01)–C X3.	–0.311 O X1. –(02)–C X2
4525	–0.108 C X3. –(07)–C X2.	–0.312 C X3. –(03)–C X2.	–0.326 O X1. –(05)–C X2

PLS	T[4]	T[5]	T[tot]
1	0.380 O X1. –(06)–C X2.	0.306 C X3. –(01)–C X2	1.571 N X3 –(02)–C X2
2	0.365 N X2 –(03)–C X3.	0.296 C X2. –(02)–C X2	1.541 N X3 –(01)–C X2
3	0.363 C X3. –(05)–C X2.	0.270 N X2 –(03)–C X3.	1.306 N X2 –(03)–C X3.
4	0.330 O X2 –(01)–N X2.	0.266 O X1. –(06)–C X2.	1.206 N X3 –(01)–C X1
5	0.323 O X2 –(03)–C X2	0.256 C X3. –(05)–C X2.	1.206 O X2 –(01)–N X2.
6	0.296 N X3 –(01)–C X1	0.232 O X1. –(02)–N X2	1.116 C X2 –(02)–C X2
7	0.288 C X3. –(01)–C X2	0.221 O X2 –(01)–N X2.	1.014 C X2 –(03)–C X2
8	0.273 C X3 –(07)–C X2.	0.220 N X3 –(01)–C X2	0.937 O X1. –(06)–C X2.
9	0.271 O X1. –(05)–C X3.	0.200 N X3 –(02)–C X2	0.920 O X2 –(03)–C X2
10	0.256 O X2 –(01)–C X3	0.195 O X1. –(05)–C X3.	0.870 C X3. –(05)–C X2.
	○	○	○
	○	○	○
4516	–0.249 O X1. –(05)–C X2	–0.192 N X3 –(02)–N X2.	–0.766 C X3. –(02)–C X1
4517	–0.249 N X3 –(04)–C X3.	–0.203 N X3 –(01)–C X3.	–0.789 N X3 –(02)–C X3.
4518	–0.256 N X3 –(03)–C X2	–0.209 N X3 –(04)–C X1	–0.812 C X2. –(01)–C X2.
4519	–0.257 N X3 –(08)–C X2.	–0.209 O X2 –(07)–C X2	–0.819 O X1. –(06)–C X1
4520	–0.278 C X3 –(07)–C X2	–0.214 N X3 –(07)–C X2.	–0.833 O X1. –(05)–C X2
4521	–0.319 C X3. –(03)–C X2.	–0.226 C X3. –(03)–C X2.	–0.861 C X3 –(07)–C X2
4522	–0.326 N X3 –(07)–C X2.	–0.247 O X1. –(06)–C X1	–0.949 N X3 –(02)–N X2.
4523	–0.345 N X3 –(02)–C X3.	–0.254 C X3. –(02)–C X1	–0.996 C X3. –(01)–C X3.
4524	–0.358 N X3 –(02)–N X2.	–0.261 C X2. –(08)–C X1	–1.217 N X3 –(01)–C X3.
4525	–0.479 N X3 –(01)–C X3.	–0.265 C X2. –(09)–C X1	–1.242 C X3. –(03)–C X2.

Fig. 5. (A) Two typical actives with example atom pairs. (B) A list of atom pairs in order of coefficient value for the 'muscarinic' set. The components are derived from sample-based PLS.

took 771 s for 50 trials. In both runs the calculation of the covariances took 120 s of the total time. The cross-validated r^2 and significance levels are:

Total components	1	2	3	4	5	6	7	8	...
Cross-validated r^2	0.129	0.214	0.285	0.320	0.329	0.297	0.252	0.209	...
Significance level (sd)	19.2	7.0	4.7	2.5	1.9	−1.0	0.2	−0.8	...

The cross-validated r^2 values indicate that five total components is optimal, while the significance levels show that component 4 is probably the last significant component.

The vectors **T**[1] to **T**[5] are listed in Fig. 5. **T**[1]–**T**[5] have many positive atom pairs in common. We interpret this as follows: compounds that have two or more of the same descriptors simultaneously are more likely to be active than expected from the trend vector prediction. A projection of the training set into the 3D space defined by the first three components is shown in Fig. 6. Again, we see that actives and inactives are better separated along **T**[tot] than along **T**[1].

Predictions

Predictions were made of the training and test sets using each method. Out of the 919 compounds in the test set, 852 could be predicted. The results of the predictions are summarized in Table 3. As expected, using more than one component results in a large improvement in the fit of the training set and a smaller, but significant, improvement in the prediction of the test set.

For large data sets, we present prediction results as a simulated screening experiment. If compounds are retrospectively tested in order of decreasing predicted activity, will the beginning of the list contain more actives relative to a randomly ordered list? Figure 7A shows this graphically for a 'prediction' of the training set. If the fit were perfect, all the actives would be at the very beginning of the list and the curves would fall on the *ideal* line. If the methods were useless, actives would accumulate approximately in proportion to their frequency in the data set, as for the *random* line. The *TV* line shows that following the predicted order of the trend vector produces actives at a much greater rate than following the random order. The five-component PLS results in a large improvement, nearly matching the ideal curve. The true predictions over the test set are shown in Fig. 7B. Here PLS gives a smaller but substantial improvement over TV. Selecting compounds according to the PLS rank would find 10–15 additional actives early in the screening.

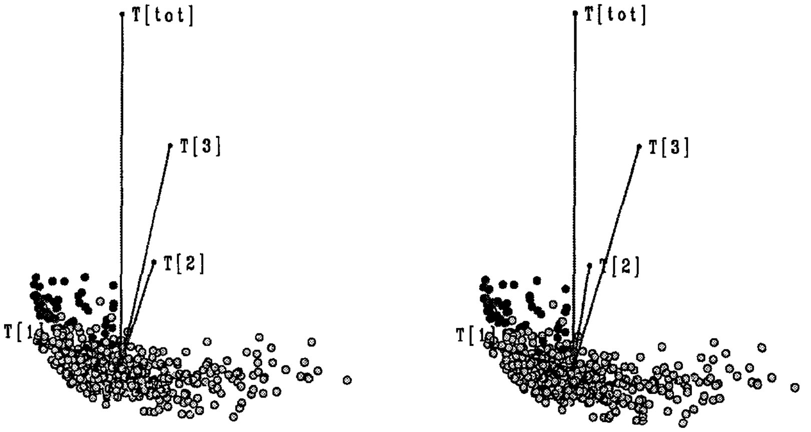


Fig. 6. The projection of each compound in the muscarinic training set in the coordinate space of the first three components T[1], T[2], T[3] of a five-component fit. The sum of the vectors T[1]–T[3] is along the direction T[tot]. 'Actives' are shown as dark circles and 'inactives' as light circles.

TABLE 3
CORRELATION BETWEEN PREDICTED AND OBSERVED ACTIVITIES FOR THE 'MUSCARINIC' SET

Method[a]	TV = ^{1}PLS	^{2}PLS	^{3}PLS	^{4}PLS	^{5}PLS
Training set	0.39	0.56	0.70	0.78	0.82
Test set	0.35	0.51	0.61	0.66	0.67

[a] Abbreviations are identical to those in Table 2.

DISCUSSION AND CONCLUSIONS

Trend vector analysis is very useful when applied with descriptors like the atom pair, especially in handling large sets of chemically diverse compounds. Yet it is computationally very simple. We have tried to improve trend vector calculations by two different methods. Trend matrix analysis applies the trend approach to an expanded set of descriptors, specifically cross-terms of the original atom pairs. In the SAMPLS approach, we use only the original descriptors, but apply a more precise linear fit. A transformation of the SAMPLS model from compound space to descriptor space is necessary to interpret the results in terms of the original descriptors. SAMPLS is the more practical of the two methods investigated. Its storage and CPU requirements are quite modest compared to trend matrix analysis, allowing it to be used on large data sets.

After examining many data sets, we find that both methods investigated produce improvements over trend vectors alone, as measured by an increase in correlation between observed and predicted activity. The rank order of predicted activity for compounds predicted to be very active or very inactive does not change much by adding correction terms. Only the predictions 'in the middle' change. Thus, more elaborate methods of predicting activity may not be necessary when selecting from a large data base a small number of compounds to be tested in a high-volume assay. (In Fig. 7B, for instance, we see that the numbers of actives found for 'TV' and '^{5}PLS' are the same at the very start of the screening.) On the other hand, when evaluating chemical structures as candidates for synthesis, it is probably worth the extra computational time to obtain better predictions.

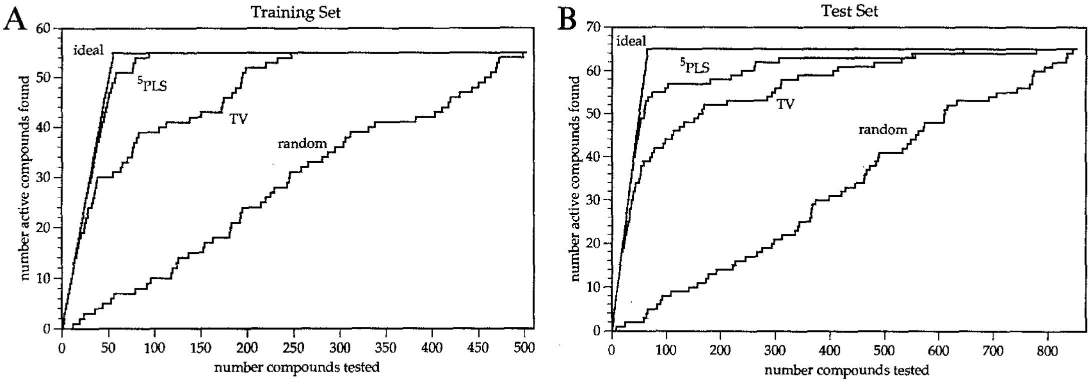


Fig. 7. Graphical representation of how quickly actives are found if the 'muscarinic' data set is 'tested' in order of decreasing predicted activity by the trend vector (TV) and by a PLS fit using five components (^{5}PLS). For comparison, the curves are shown for the idealized situations where all the actives are at the front of the data set ('ideal') or randomly scattered throughout ('random'). (A) Training set; (B) test set.

Property-based PLS is routinely used in applications like CoMFA [16]. There, the number of descriptors may be several thousands, similar to the number of topological descriptors in our second example, but the number of compounds is usually small. The SAMPLS reformulation is a much more computationally efficient way to handle this type of problem, even though it involves the extra step of transforming the results from compound space to descriptor space. When the number of descriptors and the number of compounds are both large, as in our second example, SAMPLS is the only practical approach. Our novel use of randomization as a method of determining the best number of PLS components, moreover, is a much faster alternative to leave-one-out cross-validation for large data sets. Overall, the absolute cost of running SAMPLS on large data sets with many descriptors is quite modest.

It is often stated in the PLS literature that the cross-validated r^2 is a measure of how predictive the PLS model is likely to be. In the course of this study, however, we found that this measure overestimates the number of components that will improve actual prediction. In fact, cross-validation can measure only the internal consistency of the association of descriptors with activity in the training set. For good predictions of an actual test set, the training set must also be representative of the test set. How representative the training set is depends on the level of detail at which one looks. The PLS components after the first are fitting residuals which are likely to be more sensitive to the exact composition of the training set than this first component. The randomization method for determining the number of components may be preferable, aside from its speed, because it appears more conservative than the cross-validated r^2.

AKNOWLEDGEMENTS

Dr. Ray Carhart first suggested that a matrix containing descriptor cross terms could be a useful correction to the trend vector. The form of the graphs in Fig. 7 is based on unpublished work by Dr. Norman Bauman.

REFERENCES

1 Carhart, R.E., Smith, D.H. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 25 (1985) 64.
2 Nilakantan, R., Bauman, N., Dixon, J.S. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 27 (1987) 82.
3 Klopman, G., J. Am. Chem. Soc., 106 (1984) 7315.
4 Sheridan, R.P. and Venkataraghavan, R., Acc. Chem. Res., 20 (1987) 322.
5 Nilakantan, R., Bauman, N. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 31 (1991) 527.
6 Nilakantan, R., Bauman, N. and Venkataraghavan, R., In Borowski, E. and Sugar, D. (Eds.) Proceedings of the Second Symposium on Molecular Aspects of Chemotherapy, Pergamon Press, Oxford, 1990, pp. 1–10.
7 Vityuk, N.V., Zh. Fizicheskoi Khimii, 66 (1992) 2665.
8 Bush, B.L. and Nachbar Jr., R.B., J. Comput.-Aided Mol. Design, 7 (1993) 587.
9 Wold, S., Ruhe, A., Wold, H. and Dunn III, W.J., SIAM J. Sci. Stat. Comput., 5 (1984) 735.
10 Geladi, P. and Kowalski, B.R., Anal. Chim. Acta, 185 (1986) 1.
11 Hoskuldsson, A., J. Chemometrics, 6 (1992) 307.
12 Lindgren, F., Geladi, P. and Wold, S., J. Chemometrics, 7 (1993) 45.
13 Clark, M. and Cramer III, R.D., Quant. Struct.-Act. Relatsh., 12 (1993) 137.
14 Gilligan, P.J., Cain, G.A., Christos, T.E., Cook, L., Drummond, S., Johnson, A.L., Kergaye, A.A., McElroy, J.F., Rohrbach, K.W., Schmidt, W.K. and Tam, S.W., J. Med. Chem., 35 (1993) 4344.
15 MACCS-II Drug Data Report (V. 93.1) is a product of Molecular Design Ltd, San Leandro, CA, 1993.
16 Cramer III, R.D., Patterson, D.E. and Bunce, J.E., J. Am. Chem. Soc., 110 (1988) 5959.