The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding

Robert D. Brown* and Yvonne C. Martin

Pharmaceutical Products Division, Abbott Laboratories, D47E/AP10, 100 Abbott Park Road, Abbott Park, Illinois 60064-3500

Received April 26, 1996[⊗]

We have previously studied the ability of various structural descriptors to distinguish between biologically active and inactive compounds (ref 1). This paper examines the degree to which these descriptors encode information relevant to the forces of ligand-receptor binding, namely hydrophobic, dispersion, electrostatic, steric, and hydrogen bonding interactions. This is assessed by the ability to accurately predict values for physical properties of a structure related to each of the interactions, from the known values for other structures which are shown to be structurally similar to the first by the descriptor in question. Our results suggest that the differences we observed in the ability of descriptors to separate active from inactive molecules may be explained by the degree to which they encode information relevant to ligand-receptor binding. In particular we found that the MACCS structural key descriptor implicitly contains a great deal of information relevant to each type of interaction.

INTRODUCTION

Previously[1] we reported the relative ability of several structural descriptors and clustering methods to distinguish active from inactive compounds. Across several sets of compounds that have been tested for different biological activities, we found that descriptors of 2D structure produced better separations than those of 3D structure. Best overall was a *structural key* descriptor, taken from the MACCS database software,[2] which recorded the occurrences of a number of small fragments. A typical dataset, in which 25% of structures were active, formed clusters in which an average of 85% of the compounds were active, for those clusters containing any active compounds.

We were struck by the precision of these results and hypothesized that these structural keys might contain information about those physical properties of the compounds relevant for biological activity. From the viewpoint of the target receptor in an assay, it is the intermolecular forces between it and the ligand that are important for binding and molecular recognition. These forces in turn are dependent on the physical and 3D structural properties of ligands, rather than the presence of the particular 2D fragments in the MACCS keys, and it is, therefore, these physical and 3D properties that should distinguish actives from inactives.

The binding between a ligand and receptor can be very strong and is also very specific.[3] Both strength and specificity arise from the cumulation of many weak forces between the two molecules. One such force is electrostatic interaction, the attraction or repulsion between the partial charges on atoms in the ligand and receptor. A ligand will tend to align itself within a receptor pocket to maximize the attractive electrostatic interactions whilst minimizing the repulsive interactions. Thus, electrostatic forces play an important part in molecular recognition.

Dispersion interactions arise as the result of an unequal distribution of electrons in one molecule inducing a change in the distribution of electrons in the other. Every atom in each molecule participates in dispersion interactions. These decay with the sixth power of distance between the atoms and, therefore, a strong dispersion binding will result only if there is a good shape complimentarity between the ligand and receptor, allowing a close approach of many atom pairs.

The repulsion of electron clouds around two atomic nuclei will limit their closest possible approach. This steric repulsion is large if there is a poor fit of a ligand into a receptor and so is again important in specificity.

Another important contribution to binding arises from the formation of hydrogen bonds between ligand and receptor atoms. Since hydrogen bonds are directional—they are at their strongest when the hydrogen is in direct line with the bonded and interacting atoms—hydrogen bonds again contribute to molecular recognition.

The final and energetically most important contribution to binding comes from the hydrophobic interactions that arise when two nonpolar molecules are removed from water. The increase in the entropy of the water molecules is the main contribution to the hydrophobic effect.

To understand our earlier findings, we examined the extent to which there is information encoded in any structural descriptor that relates to the interaction forces described above. Each type of interaction may be quantified in one of two ways. If available, a directly measured parameter may be used; for example, a measured log $P_{\text{octanol/water}}$ will indicate the degree of hydrophobicity of a molecule and hence its potential for hydrophobic interaction. Alternatively, inferences may be made using one or more computed properties; for example, a number of topological indices can be calculated which together give an insight into the shape of the molecules. We can assess the information content of a descriptor by the degree to which structures that are shown to be similar by that descriptor have similar properties. So, for example, if two structures have a very high similarity and very similar log $P_{\text{octanol/water}}$ values, it may be suggested that some information about hydrophobicity is encoded in the descriptors from which the similarity calculation was based.

[⊗] Abstract published in *Advance ACS Abstracts,* December 1, 1996.

S0095-2338(96)00373-3 CCC: $14.00

In this paper we report a series of experiments in which we make predictions of various properties, for each of a set of molecules, based on the mean of known values for other structurally similar molecules. Each property is chosen to probe one or more of the interaction forces discussed above. The degree to which the structural descriptor encodes information about a property is assessed by the accuracy of all the predictions made for that property.

METHODOLOGY

Descriptors and Clustering Methods. The descriptors and clustering methods examined in this study were discussed in detail in our previous paper.[1] A summary is given in tables 1 and 2. Our studies showed that across several datasets the effectiveness of the descriptors in separating biologically active and inactive molecules was the following, in decreasing order: MACCS and SSKEYS structural keys > Daylight and Unity 2D hashed fingerprints > 3D PPP pairs > Unity 3D rigid and flexible > 3D PPP triangles. The equivalent sequence for clustering methods was Wards > group average and Guénoche ≫ Jarvis−Patrick.

Property Values. A number of measured and calculated properties were chosen to give an insight into each of the factors that have been identified as important to the binding between a ligand and a receptor. These were as follows.

Measured Physical Properties. The MedChem 94b database[4] contains measured values for log *P* in octanol/water and cyclohexane/water and of pK_a. A number of the log $P_{octanol/water}$ values, known as log *P** values, are considered to be of particular accuracy by the database creators. Three datasets of structures were created from the MedChem database:

- OCT containing 8651 structures with starred octanol/water log *P* values
- CYC containing 762 structures with cyclohexane/water log *P* values
- PKA containing 8416 structures with pK_a values

The log $P_{octanol/water}$ and log $P_{cyclohexane/water}$ values are used as measures of the hydrophobicity of a molecule. In addition, the difference between the two is a measure of hydrogen bonding.[5] The dissociation constant, pK_a, will give one indication of possible electrostatic interactions.

Calculated Properties. For correlations with calculated properties we used a dataset MON, containing 1650 diverse compounds tested as monoamine oxidase inhibitors. This dataset was one of those used in our previous study. The following properties were calculated for each of these structures.

$\kappa\alpha$ descriptors,[6] $\kappa\alpha1$, $\kappa\alpha2$, and $\kappa\alpha3$, were calculated using Oxford Molecular's TSAR program.[7] These are descriptors of shape that are derived from an assumption that the shape of a molecule is a function of the number of atoms and their adjacency relationships. The $\kappa\alpha$ indices take into account the different contributions of each atom type. The three indices describe the following features of the shape of a molecule:

- $\kappa\alpha1$—the degree of complexity of the bonding pattern
- $\kappa\alpha2$—the degree of linearity or star-likeness of the bonding pattern
- $\kappa\alpha3$—the degree of branching at the center of a molecule. This will be larger for predominately linear molecules with branching at the ends.

Since dispersion interactions require that the molecule and receptor be close, a good dispersion binding will only be obtained if there is a good shape complementarity between the receptor and ligand.[3] The shape descriptors will, therefore, probe this dispersion interaction.

The flexibility index, ϕ, was calculated again using TSAR. Hall and Kier[6] note that the flexibility of a molecule is related to both its degree of linearity and the presence of branching and cycles. These factors are encoded in the $\kappa\alpha1$ and $\kappa\alpha2$ indices, and ϕ is derived from these

$$\phi = \frac{\kappa_{\alpha1}\kappa_{\alpha2}}{A}$$

where A = the number of non-hydrogen atoms.

The number of rotatable bonds in a structure also gives a simple description of its flexibility. A Daylight GCL program[8] was used to count the number of rotatable bonds in each structure. The program identifies bonds for which rotation changes the conformation of a molecule. Double and triple bonds, bonds between two atoms in the same ring, and between sp^2 carbon and the nitrogen in an amide or amidine are not considered rotatable. The following bonds are also not counted since their rotation does not change the geometry of the molecule: X−CH_3, X−C≡C, X−C≡N, X−CF_3, X−CCl_3, X−CBr_3, X−$C(Me)_3$, −C═C, X−(CH_3, CF_3, CCl_3, CBr_3, $C(Me)_3$), X−C≡(C,N), X−CO_2H, X−SO_3H, X−PO_3.

Surface area and volume were calculated using Pearlman's SAVOL2 program.[9] Molar refractivity was calculated using the CMR program from Daylight.[10] Each of these will give a measure of the steric bulk of a molecule. This, in turn, will give some indication of the fit of the molecule into the receptor site and so will provide insight into possible steric repulsions. In addition, molar refractivity is used in QSAR as a descriptor of dispersion interactions.[3]

The number of potential hydrogen-bond donors and acceptors in each molecule was counted using our in-house program 3D-features.[1] This is an expert system which uses several hundred definitions of generic and specific 2D substructures to identify points that can potentially form hydrogen bonds. The program accounts for all major tautomers and ionization states of any input structure when assigning behaviour. These values will obviously indicate the likely contribution of hydrogen bonding to the overall interactions.

Generation and Analysis of Predictions. The prediction of a property value for each compound in a dataset was made in two ways.

- **Similarity-Based Prediction.** The pairwise similarity or distance of the target compound is calculated with every other compound in the dataset. The similarity calculations can be based on any one of the descriptors listed in Table 1. The predicted value is calculated as the mean of all the values for structures that are over a given threshold similarity, or under a given threshold distance, to the target. No prediction can be made if there are no other compounds within the threshold. The more stringent the threshold, therefore, the lower the

Table 1. Descriptors Examined in the Study

	type	name	description
2D	structural key	MACCS keys[2]	153 small generic and specific fragments; number of occurrences up to 3
		SSKEYS[18]	960 small generic and specific fragments
	hashed	daylight[10]	atom types, agumented atoms and paths of 2−7; hashed to 1024 bits
		unit[17]	paths of 2−6 binned then hashed to 992 bits
3D	rigid	unity[17]	distances between pairs of heteroatoms, ring centroids and normals, and carbonyl extension points
		PPP pairs[1]	distances between pairs of hydrogen bond acceptors and donors, positive and negative charges, and hydrophobic centroids
		PPP triangles[1]	all distances between triplets of above features
	flexible	Unity[17]	all possible distances between pairs of features in Unity 3D rigid, allowing for rotation about all rotatable bonds

Table 2. Clustering Methods Examined in the Study

	type	name	description
hierarchical	agglomerative	Wards[19]	merge two clusters to minimize intracluster variance and maximize intercluster variance
		group average[19]	merges two clusters such that each cluster has a lesser average distance to the remaining members of that cluster than it does to all members of any other cluster
	divisive	Guénoche[19]	divide cluster with largest diameter into two such that larger of the two has smallest possible diameter
nonhierarchical		Jarvis−Patrick[10]	cluster two structures together if they are in each others top k most similar structures and have $I < k$ other structures in common in their top k
		enhanced Jarvis−Patrick[19]	cluster two structures together if they are in each others neighbor lists of all structures above a threshold similarity and if they have n% of the structures in those lists in common

number of predictions that can be made for a given dataset. Varying the threshold allows a number of different sets of predictions to be made for a given descriptor.

- **Cluster-Based Prediction.** The entire dataset, including the compound to be predicted, is clustered using one of the methods given in Table 2 in combination with one of the descriptors listed in table 1. The predicted value for a compound is then calculated as the mean of the values of all the other compounds in the cluster. No prediction can be made if the structure to be predicted is a singleton. Tightening the clustering conditions often results in the production of more singletons and hence allows fewer predictions to be made. In this case, a number of sets of predictions can be made by sampling the cluster hierarchies at various levels or by varying the parameters to the Jarvis−Patrick clustering.

Using a cross-validation, or leave-one-out, approach, a prediction is made for every compound in the dataset in turn. Two statistics are then calculated by comparing each predicted value to its observed (i.e., measured or calculated) value, the *product moment correlation coefficient* and the *root mean square error*.

- The product moment correlation coefficient (pmcc) indicates the correlation between the sets of observed (o) and predicted (p) values and is given by

$$\text{pmcc} = \frac{\sum(o_i - \bar{o})(p_i - \bar{p})}{\sqrt{\sum(o_i - \bar{o})^2 \sum(p_i - \bar{p})^2}}$$

This correlation coefficient ranges between 1 denoting perfect correlation, through 0 indicating no correlation, to −1 indicating perfect inverse correlation.

- Root mean square error (rmse) indicates the average size of the expected forecasting error, in the units of the measurement, over all values in the current run and is calculated by

$$\text{rmse} = \frac{\sqrt{\sum(o_i - p_i)^2}}{n}$$

Using similarity-based prediction, a number of correlations were calculated for a given dataset and descriptor by varying the threshold cutoff. For the clustering methods, results were obtained by sampling the hierarchies at different levels or by varying the parameters to the Jarvis−Patrick clustering.

RESULTS

Log $P_{\text{octanol/water}}$. Descriptors. Figure 1a shows the pmcc against the number of predictions for all structures in dataset OCT of carefully measured octanol/water log P values. Results are shown for all descriptors when predictions are made as the mean of all structures over a given Tanimoto similarity threshold. Figure 1b shows the equivalent results when predictions are made using Wards clustering. The annotations on the graphs give the rmse value for each set of predictions. The range of log P values in the dataset is −4.41 to 11.29.

Using either similarity- or cluster-based prediction, there is a clear difference between the predictive ability of the descriptors. In either case, the best predictions are obtained from the MACCS keys followed by the SSKEYS. A lower level of accuracy results when using the two hashed fingerprints from Daylight and Unity. All the 2D descriptors produce better predictions than any of the 3D ones, with the PPP pairs generally the best of these and the Unity 3D flexible fingerprints the worst. This order of predictive accuracy is exactly that found in our previous work concern-

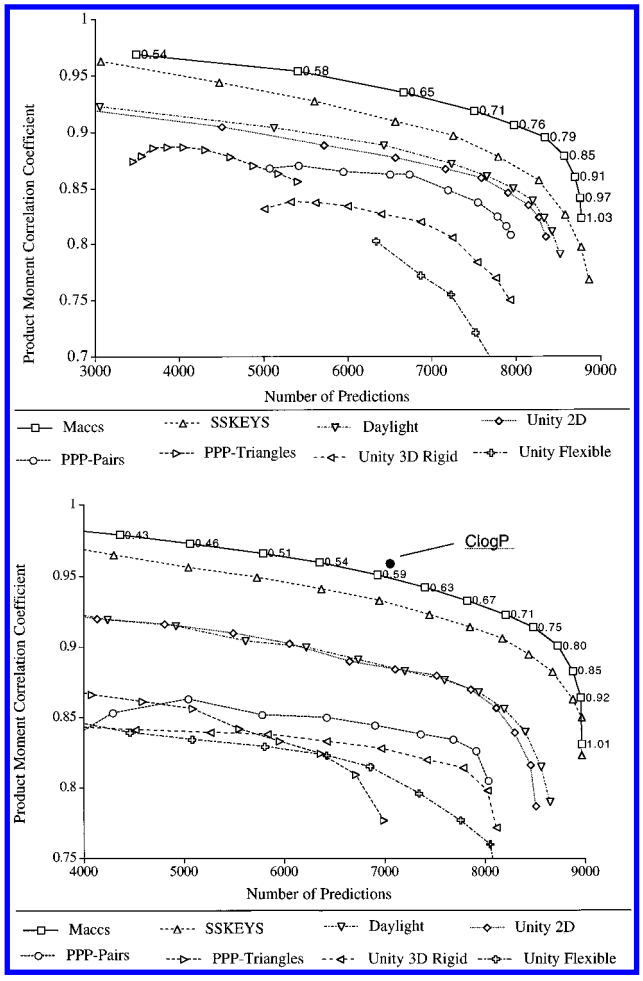

Figure 1. (a, top) PMCC vs number of predictions for predictions of log *P* from all structures in dataset OCT above a threshold Tanimoto similarity. The corresponding RMSE values are shown for the MACCS descriptors. (b, bottom) PMCC vs number of predictions for predictions of log *P* for structures in dataset OCT made from other cluster members; clusters generated using Ward's clustering. The corresponding RMSE values are shown for MACCS descriptors.

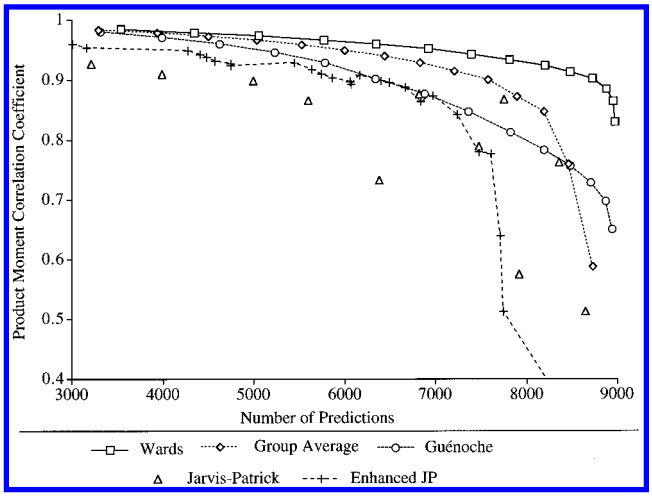

Figure 2. PMCC vs number of predictions for predictions of log *P* for structures in dataset OCT made from other cluster members; all clusters generated from MACCS key descriptors.

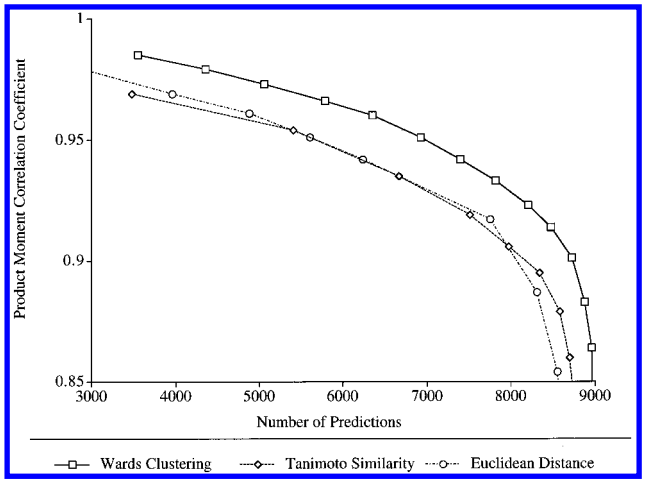

Figure 3. Comparison of similarity/distance-based predictions against cluster-based predictions of log *P* for structures in dataset OCT using MACCS keys and Ward's clustering.

ing the ability to distinguish between biologically active and inactive compounds.

The accuracy of the predictions made using the MACCS descriptors is very high. When used in conjunction with Ward's clustering, predictions can be made for approximately 8500 of the structures by using relatively "loose" clustering conditions to still achieve correlation of over 0.9 and a standard error of about 0.8 log units. This suggests that the MACCS fingerprints encode a considerable amount of information relevant to the partition coefficients of the compounds in this dataset and hence about the likely degree of the hydrophobic contribution to their interaction with a receptor.

As a comparison, the CLOGP program[10] can calculate a value of log *P* for 7054 structures from OCT with no errors or warnings. The pmcc between these values and the observed log *P** values is 0.96 with an rmse of 0.49; this point is shown in Figure 1b. Reading from graph 1b, the equivalent values for 7054 predictions (although not necessarily of the same structures) using Ward's clustering with MACCS descriptors would be a pmcc of 0.95 and rmse of 0.60. Note that predictions considerably more accurate than CLOGP can be made if a higher similarity cutoff than this is applied. This requires only that sufficiently similar structures to the one being predicted exist in the dataset.

Clustering Methods. Figure 2 shows the differences in predictive ability of the various clustering methods using MACCS keys. Ward's clustering produces the most accurate predictions, followed by group average and then Guénoche. Standard Jarvis−Patrick is, for the most part, the worst method. In a few cases, under particular clustering conditions, it is almost as good as Guénoche. However, the method is more erratic. Two different sets of Jarvis−Patrick clusters, each of which allow approximately 8000 predictions to be made, give correlation coefficients of around 0.6 and 0.9. The enhanced Jarvis−Patrick produces results similar to the standard method with loose clustering conditions but outperforms the standard method when the conditions are more strict. The trend in the predictive ability of the clustering methods once again reflects that found in the analysis of the biological activity data.

The predictions produced from the MACCS keys by Tanimoto similarity or Euclidean distance alone are compared to those produced using Ward's clustering in Figure 3. (Ward's clustering makes use of Euclidean distance.) There is little difference between the predictions made using the distance measure instead of the similarity.

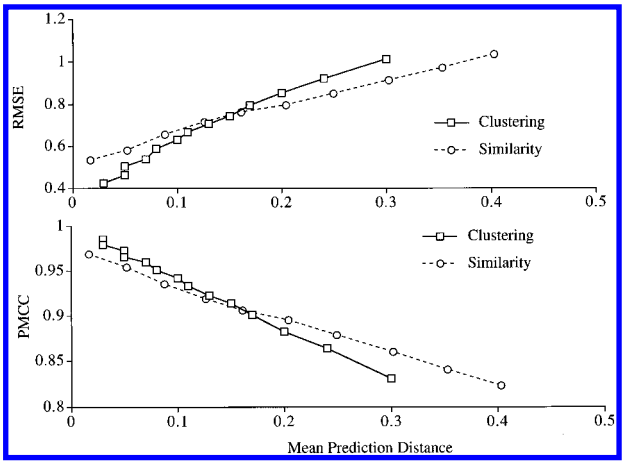

Figure 4. Mean prediction distance against (a) rmse and (b) pmcc for similarity and cluster-based predictions of log *P* for structures in dataset OCT using MACCS keys and Ward's clustering.

Figure 3 also shows that for a given number of predictions, greater accuracy can be obtained using Ward's clustering than using pairwise similarity alone. The *mean prediction distance* for a single set of predictions may be calculated by first recording the mean distance from a structure for which a prediction is being made to all other structures, either in a cluster or above a threshold, on which that prediction is based. Taking the mean value of this over all structures in a dataset for which predictions are being made gives the mean prediction distance. Figure 4a shows a plot of rmse against this mean prediction distance for both similarity- and cluster-based predictions of the structures in OCT. Figure 4b shows a corresponding plot for pmcc against mean prediction distance. These plots show that with a tight threshold and consequently low mean prediction distance, cluster-based predictions are the more accurate. However, as the conditions are relaxed the quality of the clustering predictions deteriorates more rapidly than those based on similarity.

Using the following procedure, it can be shown that there is a statistically significant difference between the slopes of the two lines in Figure 4b. The dataset was split into five parts by random selection. Equivalent calculations to those described above were repeated to give a graph of the type shown in Figure 4b but with five lines for each type of prediction. An equation for each line was produced using linear regression. This produces five pairs of slopes, each pair corresponding to one subset of the data and having one slope for cluster-based prediction and the other for similarity. These may be compared using a paired t-test[11] on the differences between the slopes. This produces a t-statistic of 3.75 which is significant at the 95% level.

Predictions of Other Properties. Using graphs of the type shown in Figure 1, it is possible to obtain an estimate for the pmcc and rmse for a given property for a fixed number of predictions and hence compare the different descriptors, clustering methods, and prediction methods directly.

The remaining results are presented in summary only, by examining only one data point per series. The graphs from which these data points were taken show similar patterns to those presented in Figure 1; in particular there is generally an even gap between any two distributions in these graphs over a wide range of numbers of predictions. Results from a subset of descriptors are presented: MACCS keys are taken as representative of the structural keys, Unity 2D fingerprints as representative of the hashed 2D fingerprints, PPP pairs as representative of the 3D rigid fingerprints, and Unity 3D flexible fingerprints included as the only example of its type. In all cases Ward's clustering produced the most accurate predictions, and so all clustering results use this method. Furthermore, there was found to be little advantage to using Tanimoto similarity or Euclidean distance; therefore, all similarity-based predictions are made using the former.

Table 3 shows the accuracy of prediction for measured log $P_{\text{cyclohexane/water}}$ at a level of similarity or clustering for which predictions can be made for 650 of the 762 structures in CYC. Table 4 shows equivalent results for 7000 predictions of the 8416 measured pK_a values. Tables 5−14 show comparable results for 1400 predictions of the calculated properties $\kappa\alpha1$, $\kappa\alpha2$, $\kappa\alpha3$, ϕ, number of rotatable bonds, surface area and volume, molar refractivity, and number of potential hydrogen bond donors and acceptors, respectively, for the 1650 structures in dataset MON.

Table 3. Predictions of log $P_{\text{cyclohexane}}$ for 650 Structures from Dataset CYC[a]

	similarity prediction		cluster prediction	
descriptor	pmcc	rmse	pmcc	rmse
MACCS keys	0.83	0.87	0.86	0.82
Unity 2D	0.79	0.95	0.84	0.90
PPP pairs	0.75	1.0	0.77	0.96
Unity flex	0.69	1.1	0.71	1.1

[a] Range of values in the dataset is −4.00 to 5.29.

Table 4. Predictions of pK_a for 7000 Structures from Dataset PKA[a]

	similarity prediction		cluster prediction	
descriptor	pmcc	rmse	pmcc	rmse
MACCS keys	0.81	1.8	0.87	1.6
Unity 2D	0.81	1.8	0.87	1.6
PPP pairs	0.72	2.3	0.75	2.2
Unity flex	0.69	2.4	0.73	2.3

[a] Range of values in the dataset is −9.41 to 14.97.

Table 5. Predictions of $\kappa\alpha1$ for 1400 Structures from Dataset MON[a]

	similarity prediction		cluster prediction	
descriptor	pmcc	rmse	pmcc	rmse
MACCS keys	0.79	2.9	0.78	3.0
Unity 2D	0.63	3.7	0.66	3.7
PPP pairs	0.66	3.7	0.63	3.8
Unity flex	0.56	4.2	0.56	4.2

[a] Range of values in the dataset is 2.3−42.5.

Table 6. Predictions of $\kappa\alpha2$ for 1400 Structures from Dataset MON[a]

	similarity prediction		cluster prediction	
descriptor	pmcc	rmse	pmcc	rmse
MACCS keys	0.79	1.8	0.80	1.8
Unity 2D	0.61	2.4	0.69	2.2
PPP pairs	0.72	2.2	0.64	2.4
Unity flex	0.53	2.6	0.55	2.6

[a] Range of values in the data set is 1.0−27.7.

Table 7. Predictions of $\kappa\alpha 3$ for 1400 Structures from Dataset MON[a]

	similarity prediction		cluster prediction	
descriptor	pmcc	rmse	pmcc	rmse
MACCS keys	0.78	2.3	0.77	2.3
Unity 2D	0.70	2.5	0.65	2.8
PPP pairs	0.65	2.8	0.59	3.0
Unity flex	0.47	3.3	0.48	3.4

[a] Range of values in the dataset is 0.4−47.2.

Table 8. Predictions of ϕ for 1400 Structures from Dataset MON[a]

	similarity prediction		cluster prediction	
descriptor	pmcc	rmse	pmcc	rmse
MACCS keys	0.82	1.7	0.82	1.7
Unity 2D	0.68	2.1	0.67	2.2
PPP pairs	0.71	2.1	0.68	2.1
Unity flex	0.50	2.7	0.52	2.6

[a] Range of values in the dataset is 0.6−29.0.

Table 9. Predictions of Number of Rotatable Bonds for 1400 Structures from Dataset MON[a]

	similarity prediction		cluster prediction	
descriptor	pmcc	rmse	pmcc	rmse
MACCS keys	0.78	2.5	0.80	2.4
Unity 2D	0.55	3.3	0.63	3.1
PPP pairs	0.76	2.6	0.72	2.7
Unity flex	0.52	3.4	0.55	3.3

[a] Range of values in the dataset is 0−34.

Table 10. Predictions of Surface Area for 1400 Structures from Dataset MON[a]

	similarity prediction		cluster prediction	
descriptor	pmcc	rmse	pmcc	rmse
MACCS keys	0.78	55	0.80	54
Unity 2D	0.61	70	0.67	68
PPP pairs	0.77	57	0.74	59
Unity flex	0.61	70	0.61	72

[a] Range of values in the dataset is 59−927.

Table 11. Predictions of Volume for 1400 Structures from Dataset MON[a]

	similarity prediction		cluster prediction	
descriptor	pmcc	rmse	pmcc	rmse
MACCS keys	0.80	47	0.80	47
Unity 2D	0.66	58	0.70	57
PPP pairs	0.77	57	0.74	59
Unity flex	0.65	60	0.63	61

[a] Range of values in the dataset is 35−767.

In all cases except the hydrogen bonding counts, and for either cluster or similarity-based prediction, the MACCS keys are shown to produce the most accurate predictions. In all cases the Unity 3D flexible fingerprints produce the poorest predictions. For the measured properties, log $P_{\text{cyclohexane/water}}$, and p$K_a$, the Unity 2D fingerprints do not perform as well as the MACCS keys but perform better than the PPP pairs. For some of the calculated properties, notably number of rotatable bonds, surface area, volume and molar refractivity, the predictive order of the PPP pairs and Unity 2D fingerprints is reversed; for the shape indices the two are

Table 12. Predictions of Molar Refractivity for 1400 Structures from Dataset MON[a]

	similarity prediction		cluster prediction	
descriptor	pmcc	rmse	pmcc	rmse
MACCS keys	0.79	1.5	0.77	1.5
Unity 2D	0.67	1.8	0.68	1.7
PPP pairs	0.79	1.5	0.75	1.5
Unity flex	0.69	1.8	0.67	1.8

[a] Range of values in the dataset is 0.9−22.7.

Table 13. Predictions of Number of Hydrogen Bond Acceptors for 1400 Structures from Dataset MON[a]

	similarity prediction		cluster prediction	
descriptor	pmcc	rmse	pmcc	rmse
MACCS keys	0.83	0.9	0.85	0.8
Unity 2D	0.74	1.1	0.76	1.1
PPP pairs	0.90	0.7	0.87	0.8
Unity flex	0.69	1.2	0.74	1.1

[a] Range of values in the dataset is 0−14.

Table 14. Predictions of Number of Hydrogen Bond Donors for 1400 Structures from Dataset MON[a]

	similarity prediction		cluster prediction	
descriptor	pmcc	rmse	pmcc	rmse
MACCS keys	0.79	0.8	0.80	0.8
Unity 2D	0.76	0.9	0.76	0.9
PPP pairs	0.91	0.6	0.89	0.6
Unity flex	0.50	1.2	0.59	1.1

[a] Range of values in the dataset is 0−10.

similar. In one of the biological activity datasets we examined in our earlier work,[1] there was no clear difference between the PPP pairs and 2D hashed fingerprints, although in the others the latter was the better of the two.

Predictions of the number of potential hydrogen bond donors and acceptors are the only cases in which MACCS structural keys were not most accurate. However, the same program that was used to calculate the number of each was also used to produce the PPP pair fingerprints, in which separate regions of the fingerprint encoded each donor−donor, donor−acceptor, and acceptor−acceptor distance. It is hardly surprising, therefore, that this fingerprint produces the best predictions in this case. Furthermore, the predictions made using the MACCS fingerprints are as accurate as the predictions of several other properties using this descriptor.

Whilst not as high as the correlations observed with log $P_{\text{octanol/water}}$, the correlations for MACCS keys, and optionally Ward's clustering, with the two other measured properties, log $P_{\text{cyclohexane/water}}$ and pK_a, are also very high, being close to 0.9 whilst still making predictions for most of the dataset. The correlations with the various calculated properties are somewhat lower but are still around 0.8. This is remarkable since the clusters at the appropriate level of the hierarchy are by no means very small and "narrow" and the similarity cutoff level for similarity-based prediction is not particularly high, being typically between 0.7 and 0.8.

Klopman and Fercu[12] have described a fragment based expert system (Multi-CASE) for the prediction of pK_a. Using a training set of approximately 2500 acids they predicted pK_as for a further 200 acids with a correlation of 0.82 and

Table 15. Correlation Matrix for the Ten Calculated Properties

	ka1	ka2	ka3	flex	rot bonds	sa	vol	cmr	acceptors	donors
ka1	1.00	0.85	0.56	0.79	0.75	**0.94**	**0.94**	**0.91**	0.03	0.41
ka2		1.00	0.77	**0.97**	0.88	0.85	0.81	0.76	−0.02	0.31
ka3			1.00	0.86	0.77	0.55	0.47	0.38	−0.04	0.13
flex				1.00	0.88	0.76	0.70	0.63	−0.03	0.28
rot bonds				1.00	0.88	0.78	0.72	0.62	−0.06	0.28
sa					1.00	1.00	**0.99**	**0.95**	−0.02	0.29
vol						1.00	**0.98**	−0.03	0.28	
cmr							1.00	−0.02	0.23	
acceptors								1.00	0.28	
donors									1.00	

Table 16. Eigenvalues of the Correlation Matrix

	1	2	3	4	5	6	7	8	9	10
eigenvalue	6.60	1.27	1.07	0.66	0.18	0.13	0.06	0.02	0.01	0.00
difference	5.33	0.21	0.41	0.49	0.05	0.07	0.03	0.02	0.00	
proportion	0.66	0.13	0.11	0.07	0.02	0.01	0.01	0.00	0.00	0.00
cumulative	0.66	0.79	0.89	0.96	0.98	0.99	1.00	1.00	1.00	1.00

an rmse of 1.58. By comparison, the MACCS/Wards cluster-based predictions of 7000 acids and bases achieved a correlation of 0.87 and an rmse of 1.6.

If the various physical properties for each structure are highly correlated, then we have merely shown that the descriptors are encoding information relevant to only a very few independent factors. It is, therefore, interesting to establish the degree of independence between the physical properties to see if this is indeed the case. This we have done first by calculating the pairwise correlations between each pair of calculated properties for a given structure and, secondly, by conducting a principal components analysis on the ten calculated properties to establish the minimum number of independent variables required to represent those properties. Note that it was not possible to include log *P* and pK_a in this analysis since these are for different sets of structures than the calculated properties.

Table 15 shows the pairwise correlation matrix for the ten properties over the structures in the MAO dataset. There are a number of high correlations, for example, between surface area and volume and between $\kappa\alpha 1$ and surface area or volume or cmr; pairwise correlations of over 0.9 are highlighted. However, the table shows that not all the properties are directly correlated with each other. The eigenvalues of a principal components analysis are shown in Table 16. This suggests that four, or possibly five, principal components are needed to represent the ten variables. Together these analyses show that the descriptors are not simply encoding information about a single variable with which the properties are all correlated but rather are encoding information relevant to a number of independent pieces of information.

For the three measured properties there is a clear advantage to the cluster-based prediction over similarity-based prediction for a given descriptor. For the calculated properties there is typically little advantage to one method over the other. In all cases predictions made using clusters are based, on average, on more similar structures than are predictions based on similarity, although the gap between the two is considerably greater for all measured properties.

Since the calculated and measured properties are for different sets of structures, it is necessary to establish that the difference between cluster and similarity based prediction is not due to this factor. Figure 5 shows the cluster and similarity based predictions of the calculated surface area for the structures in the log *P** dataset, OCT, together with the predictions of log *P*. For the same set of structures, the two methods are performing equally for the calculated property which suggests that it is not the difference in the sets of structures which is causing the difference between the two types of predictions. However, we can offer no clear explanation of why clustering should be favored for prediction of the measured properties but not the calculated ones.

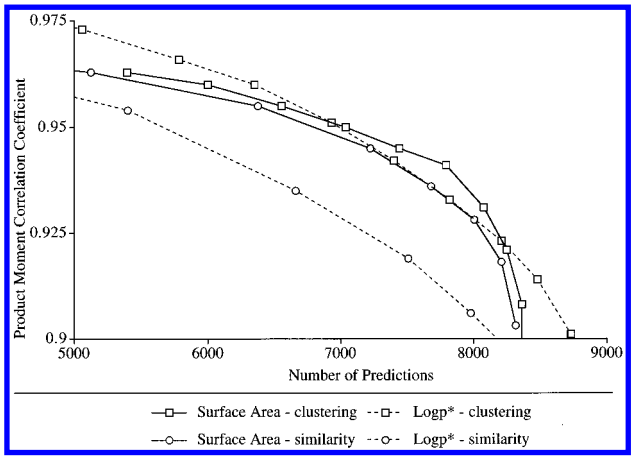

Figure 5. Predictions of surface area and log *P* for structures in dataset OCT made using both similarity- and cluster-based prediction.

Figure 5 also suggests that the lower accuracy of prediction observed for the calculated properties on the MAO dataset, compared to the measured properties on the MedChem94 datasets is a result of the differences in the set of structures rather than the types of properties. The predictions of surface area and log *P* for dataset OCT are approximately of equal accuracy. Instead the difference may be a result of the difference in the diversities of the structures in the two datasets.

One simple measure of diversity may be obtained from the Ward's clustering method.[13] At each level of the hierarchy a cluster separation statistic is obtained which relates to the distance between the two clusters to be merged at that level. The equivalent partition in the hierarchies of two different sets of clusters may be found by extracting clusters with an equal separation. At a given hierarchy level, a diverse dataset will have formed more clusters, relative to

Table 17. Diversities of the Four Datasets at Cluster Separation 10.0

dataset	diversity	dataset	diversity
OCT	0.29	PKA	0.30
CYC	0.29	MON	0.49

the total number of structures, than a less diverse dataset which will instead have formed fewer clusters each containing more structures. A simple measure of the diversity of a dataset may therefore be obtained as the number of clusters at a given level divided by the number of structures. A completely diverse dataset will score 1.0 on this scale and a completely homogeneous dataset the inverse of the total number of structures. Table 17 shows the diversities, by this measure, of the four datasets. MON is shown to be much more diverse than the other datasets, which all derive from the same database. For MON, therefore, there are likely to be fewer structures in each cluster, or above each similarity threshold, on which to base a prediction, and so the predictions might be expected to be slightly less accurate.

In summary, the diversity of the structures in the dataset appears to influence the absolute accuracy of the predictions made but not whether cluster-based prediction will be more effective than similarity-based prediction.

DISCUSSION

The results suggest that the trends that we observed in the ability of particular descriptors to distinguish biologically active from inactive molecules may be explained by the degree to which each descriptor encodes information about the properties of a ligand relevant to receptor binding. Our previous observation that 2D structural keys perform better than 2D hashed fingerprints, which perform better than any of the 3D descriptors we examined, is directly mirrored by the order of the descriptors' predictive ability for all of the properties examined in this study. For biological activity we found that Ward's clustering outperformed the group average and Guénoche methods which were, in turn, better than Jarvis−Patrick. This trend is once again exactly that found when examining the accuracy of the predictions of physical properties made using each of these clustering methods.

Over a wide variety of different types of property, the best predictions were obtained using the MACCS structural keys. These were found to encode the greatest amount of information about hydrophobicity, electrostatics, sterics, dispersion interactions, and hydrogen bonding. We previously observed that the MACCS keys functioned best over several different types of biological activity. Our current results suggest that this might apply generally over many other types of activity, since the keys best cover most of the important factors for ligand-receptor binding and molecular recognition.

The MACCS keys are of a variety of types including atom counts, ring types and counts, augmented atoms, and short linear sequences. They were selected for optimum screenout during substructure search, not for similarity searching to reflect similarities in biological or physical properties. Whilst it is not easy to explain why the keys should be so well suited to this task, it appears that these combinations of fragments must be sufficient to characterize much of the 3D structure of the molecules. In particular the augmented atoms recording the various branching patterns together with the linear sequences recording 2D distances between atoms probably contribute to this ability. The 2D descriptors also have the advantage that any 3D information they are implicitly encoding will be conformation independent.

Other approaches have been described to ensure that similarity and clustering methods take account of the various factors important in ligand-receptor binding. Downs *et al.*[14] have reported a method for similarity and clustering based directly on physical property descriptors rather than structural descriptors. In this way properties including log *P*, molar refractivity, and volume are included directly. One disadvantage to this is that all calculated properties must be available for every structure in the dataset. Furthermore, a number of properties encoding electronic and electrostatic information were included.[15] These were calculated using MOPAC, which could potentially be a time consuming process; processing times of 30 s to 1 min per structure are not unusual in our experience.

An alternative approach by Martin *et al.*[16] is to produce a composite descriptor which contains elements of a structural descriptor together with some physical property descriptors. Lipophilicity is represented as a calculated octanol/water partition coefficient, taken from a number of commercially available programs. Side chain shape, flexibility, and branching are encoded using topological indices from Molconn-X. Chemical functionality descriptors are contained in the Daylight structural fingerprints. Receptor recognition descriptors encode the acidity, basicity, likely hydrogen bonding behavior, and aromaticity of atoms in the molecule. Once again, this technique relies on being able to calculate all values for all structures. Principal components analysis and multidimensional scaling are used to reduce the information to a small number of variables which are then used to calculate similarity coefficients.

Our results suggest that by making an appropriate choice of structural descriptor, and of clustering method where applicable, the ligand-receptor binding forces can be accounted for without having to explicitly code them in the descriptor. There are two major advantages to this. One is that it is not necessary to be able to calculate all property values for every structure. The MACCS keys seem to account well for the degree of hydrophobicity for all structures in the set of structures with highly accurate measured log *P* values. By contrast, CLOGP was unable to produce a high confidence estimate for well over 1000 of those structures and another calculation program would have had to be used to account for these. A second advantage of using only fragment descriptors is speed and simplicity. The MACCS keys, when produced using the Tripos fingerprinting utility[17] require of the order of 0.5 s CPU per structure to generate on a fast workstation.

CONCLUSION

We have shown that there is a direct relationship between the amount of information relevant to ligand-receptor binding which a descriptor encodes and that particular descriptor's ability to distinguish between biologically active and inactive structures. The accuracy of the predictions of properties such as log *P* produced using the MACCS structural keys were very high and comparable to the predictions of accepted

methods such as CLOGP. We have not set out to create a new method for property prediction; however, in the light of the results presented here, this may merit further investigation.

ACKNOWLEDGMENT

We would like to thank MDL Information Systems for access to the SSKEYS gateway, Barnard Chemical Information Ltd. for access to the enhanced Jarvis−Patrick program, and Professor Peter Willett for his helpful comments on a draft of this paper.

REFERENCES AND NOTES

(1) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(2) *Maccs II*; Molecular Design Ltd.: 14600 Catalina St, San Leandro, CA 94577. (510)-895-1313.
(3) Martin, Y. C. In *Modern Drug Research. Paths to Better and Safer Drugs*; Martin, Y. C., Kutter, E., Austel, V., Eds.; Marcel Dekker: New York, 1989.
(4) *MedChem 94b*; BioByte Corp.: 645 N. College Ave, Claremont, CA. (909) 621 8452.
(5) Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.
(6) Hall, L. H.; Kier, L. B. In *Reviews in Computation Chemistry, Volume 2*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1991.
(7) *TSAR v2.2*; Oxford Molecular, Ltd.: Oxford Science Park, Oxford OX4 4GA, UK. (+44) 1865 784600. products@oxmol.co.uk.
(8) *MedChem ver 3.54*; Daylight Chemical Information, Inc.: 27401 Los Altos, Suite #370, Mission Viejo, CA 92691. 714-367-9990, info@daylight.com.
(9) Pearlman, R. *SAVOL 2*; College of Pharmacy, The University of Texas at Austin: Austin, TX. pearlman@vax.phr.utexas.edu.
(10) *Daylight Chemical Information Software, ver4.41*; Daylight Chemical Information, Inc.: 27401 Los Altos, Suite #370, Mission Viejo, CA 92691. 714-367-9990, info@daylight.com.
(11) *SAS*; SAS Institute Inc.: Cary, NC.
(12) Klopman, G.; Fercu, D. Application of the Multiple Computer Automated Structure Evaluation Methodology to a Quantitative Structure-Activity Relationship Study. *J. Comput. Chem.* **1994**, *15*, 1041−1050.
(13) Brown, R. D.; Bures, M. G.; Martin, Y. C. Similarity and Cluster Analysis Applied to Molecular Diversity. 209th Meeting of the American Chemical Society, Anaheim, CA; American Chemical Society: Washington, DC.
(14) Downs, G. M.; Willett, P.; Fisanick, W. Similarity Searching and Clustering of Chemical Structure Databases Using Molecular Property Data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094−1102.
(15) Fisanick, W.; Cross, K. P.; Rusinko III, A. Characteristics of Computer-Generated 3D and Related Molecular Property Data for CAS Registry Substances. *Tetrahedron Comput. Methodol.* **1990**, *3*, 635−652.
(16) Martin, E. J.; Blaney, J. M.; Siani, M. S.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.
(17) *Unity Chemical Information Software ver 2.3*. Tripos Associates: 1699 S Hanley Road, Suite 303, St Louis, MO 63144. 1-800-323-2960, support@tripos.com.
(18) *SSKEYS Gateway*; MDL Information Systems, Inc.: 14600 Catalina St, San Leandro, CA 94577. (510)-895-1313.
(19) *BCI Clustering Package, versions 2.5 & 3.0*; Barnard Chemical Information, Ltd.: 46 Uppergate Road, Sheffield, S6 6BX UK. (+44)-114-233-3170, barnard@bci1.demon.co.uk.

CI960373C