The Utilization of Reduced Dimensional Representations of Molecular Structure for Rapid Molecular Similarity Calculations

Daniel D. Robinson, Thomas W. Barlow, and W. Graham Richards*

Physical and Theoretical Chemistry Laboratory, Oxford University, South Parks Road,
Oxford, OX1 3QZ, U.K.

Received March 14, 1997[⊗]

The availability of two-dimensional representations of molecules which retain structural information permits the application of techniques from digital image processing to be applied to molecular similarity. Here three such approaches, invariant moments, radial integration and radial scanning, are investigated. They overcome the time consuming adjustment of the relative orientations of the molecules to be compared so as to optimize the similarity and can speed up the overall calculations by many orders of magnitude, hence offering a technique for dealing with the enormous numbers of compounds being generated by high throughput synthesis and combinatorial chemistry. These techniques are applied to the well-known set of steroids which are frequently used as a benchmark set for QSAR studies.

INTRODUCTION

We have shown in previous papers how low error 2D representations of molecules can be generated by a nonlinear mapping (NLM) algorithm. These representations retain most of the three-dimensional distance geometry information and offer the possibility of greatly accelerating calculations of molecular similarity by permitting the application of techniques developed in image processing, particularly optical character recognition. The importance of this massive increase in speed of computation derives from the huge growth in the number of compounds which are being generated from techniques such as combinatorial chemistry and robotic high throughput synthesis.

MOLECULAR SIMILARITY

The most widely used metric for similarity calculation is the Carbo[1] index which in three dimensions can be written as

$$C_{\mathrm{AB}} = \frac{\int\int\int \rho_{\mathrm{A}}(x,y,z)\rho_{\mathrm{B}}(x,y,z)\mathrm{d}x\mathrm{d}y\mathrm{d}z}{\sqrt{\int\int\int \rho_{\mathrm{A}}^{2}(x,y,z)\mathrm{d}x\mathrm{d}y\mathrm{d}z\int\int\int \rho_{\mathrm{B}}^{2}(x,y,z)\mathrm{d}x\mathrm{d}y\mathrm{d}z}}$$

where $\rho(x,y,z)$ is a property of the molecules A and B, for example, electron density, electrostatic potential,[2] or shape.[3] Clearly the triple integrals in this equation make it a time consuming expression to evaluate. In two dimensions things are a little better as the Carbo index can be written in a form containing only two-dimensional integrals

$$C_{\mathrm{AB,2D}} = \frac{\int\int \rho_{\mathrm{A,2D}}(x,y)\rho_{\mathrm{B,2D}}(x,y)\mathrm{d}x\mathrm{d}y}{\sqrt{\int\int \rho_{\mathrm{A,2D}}^{2}(x,y)\mathrm{d}x\mathrm{d}y\int\int \rho_{\mathrm{B,2D}}^{2}(x,y)\mathrm{d}x\mathrm{d}y}}$$

where the subscript 2D denotes that the properties have been projected onto the two-dimensional representation.

Now that we have the equation in a form which we can easily represent on a sheet of paper let us examine what the Carbo index is doing. For our example we shall consider the metric of shape, where the atoms are represented as circles whose radii are equal to the van der Waals radius of the atom.

Let us project our two-dimensional representations of our two molecules onto grids as shown in Figure 1. Evaluating the index for our two molecules simply involves counting up those grid elements which contain an atom in both molecules. The more places each molecule has an atom in common, the more similar they are. The denominator in the index forces it to return a number between 0 and 1.

This seems relatively straightforward. We simply perform the nonlinear map of the structure, draw the resulting molecule on some "virtual" graph paper, and count up the number of times an atom occurs at the same position. Unfortunately relative to each other, our two-dimensional representations can exist with any arbitrary offset and any arbitrary rotation, hence forming a three-dimensional "Relative Configuration Space" (RCS). Any movement in RCS of the two-dimensional representations from the optimum will lower the similarity reported by the index.

For example, consider Figure 2. In this case the molecules are rotated 90° relative to each other and offset. Only the dark grey areas will be counted by the index as being similar, and the wrong value would be obtained for the molecular similarity.

From this we can see that any calculation using the Carbo index is liable to require a complete search of Relative Configuration Space. Even with a two-dimensional representation this will take some considerable time, but there are other metrics we can use to speed up the similarity calculation.

These come from the area of Digital Image Processing.[4,5] These algorithms, which have been highly developed over many years of research, are designed to compare quickly a two-dimensional image taken from the outside world with a database of previously captured images. The comparison is then used to decide to which "class" the captured image belongs. A particular instance of such a process would be

[⊗] Abstract published in *Advance ACS Abstracts,* August 1, 1997.

S0095-2338(97)00423-X CCC: $14.00

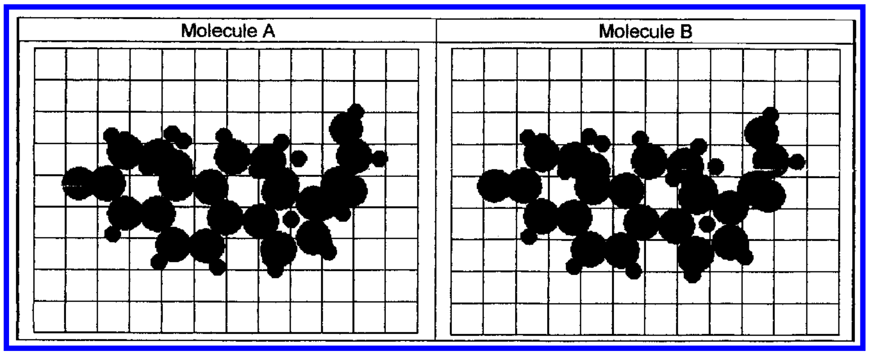


Figure 1. Projection of two steroid molecules on their calculation surface.

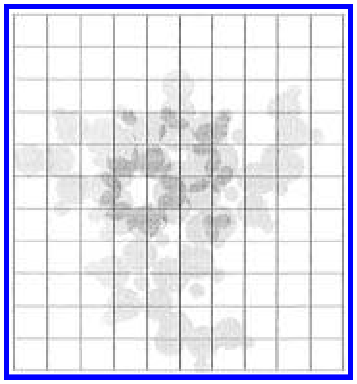


Figure 2. The effect of nonoptimal alignment.

optical character recognition (OCR) where the computer is required to read in a page of text, in the form of an image from a page scanner, and convert this into a machine editable format. In this case the database of previous images would be the entire character set in a particular font, whilst the captured image would be the separate letters which make up the complete page. The letters would clearly be sorted into "classes" according to which letter of the alphabet they were, i.e., upper or lower case "A", etc. Considering that a scanned image may contain 4096 × 4096 pixels and that the required processing rate is of the order of one page per min, it is clear that the image processing algorithms are vastly more efficient than the traditional similarity indices used in chemistry.

We shall consider three techniques to evaluate their relative performance and suitability for chemical similarity measurements and database searching.

INVARIANT MOMENTS

This method of pattern recognition was introduced by Hu[6] in 1965. The method is based upon a statistical analysis of the distribution and value of the pixels which make up the image to be compared. This distribution takes the form of the following equation

$$m_{\mathrm{p,q}} = \int\int x^{\mathrm{p}} y^{\mathrm{q}} \rho(x,y)\mathrm{d}x\mathrm{d}y \quad \text{for continuous data}$$

or

$$m_{\mathrm{p,q}} = \sum\sum x^{\mathrm{p}} y^{\mathrm{q}} \rho(x,y) \quad \text{for discrete data}$$

From a uniqueness theorem due to Papoulis,[7] provided that $\rho(x,y)$ is continuous and has nonzero values only in a finite part of the x,y plane, moments of all orders exist, and the sequence of moments $m_{\mathrm{p,q}}$ is uniquely determined by $\rho(x,y)$. Conversely it can be shown that the infinite sequence of $m_{\mathrm{p,q}}$ uniquely determines $\rho(x,y)$.

It is immediately apparent that the above equations suffer from the same problem as the traditional similarity indices, in that they are sensitive to translation and rotation. However, in his paper Hu shows how these can be dealt with.

GAINING TRANSLATIONAL INSENSITIVITY

Inherent in the equations for generating the moments is the assumption that the property $\rho(x,y)$ is centered upon the calculation coordinates. This may not be the case. However, let us define two parameters as follows:

$$\bar{x} = \frac{m_{1,0}}{m_{0,0}}$$

$$\bar{y} = \frac{m_{0,1}}{m_{0,0}}$$

These two parameters clearly give the center of the property $\rho(x,y)$ in the calculation coordinate system. This has shown to be consistently positioned for all molecules which can be reasonably compared. We can now define the central moment $\mu_{\mathrm{p,q}}$ by the following equation

$$\mu_{\mathrm{p,q}} = \int\int (x - \bar{x})^{\mathrm{p}} (y - \bar{y})^{\mathrm{q}} \rho(x,y)\mathrm{d}x\mathrm{d}y \quad \text{for continuous data}$$

or

$$\mu_{\mathrm{p,q}} = \sum\sum (x - \bar{x})^{\mathrm{p}} (y - \bar{y})^{\mathrm{q}} \rho(x,y) \quad \text{for discrete data}$$

These central moments have the desired invariance to translation.

GAINING INVARIANCE TO ROTATION

This is a much more complex operation than the above process, and only the outline of the operations contained within Hu's paper will be discussed.

Aligning a structure along its principal axes enables us to remove any unwanted rotation. In the case of the moments Hu utilizes this property of the principal axes to rotate $\rho(x,y)$ so that the property's distribution is aligned along the calculation axes. This is done by forming a variety of combinations of the three second order central moments and the four third order central moments. In his paper Hu explains the logic behind these combinations, which are able to distinguish between structures which exhibit mirror and rotational symmetry:

Hu maintains that these seven invariant central moments are all that is required to gain a high degree of sensitivity in pattern recognition. Indeed in his own paper he used only

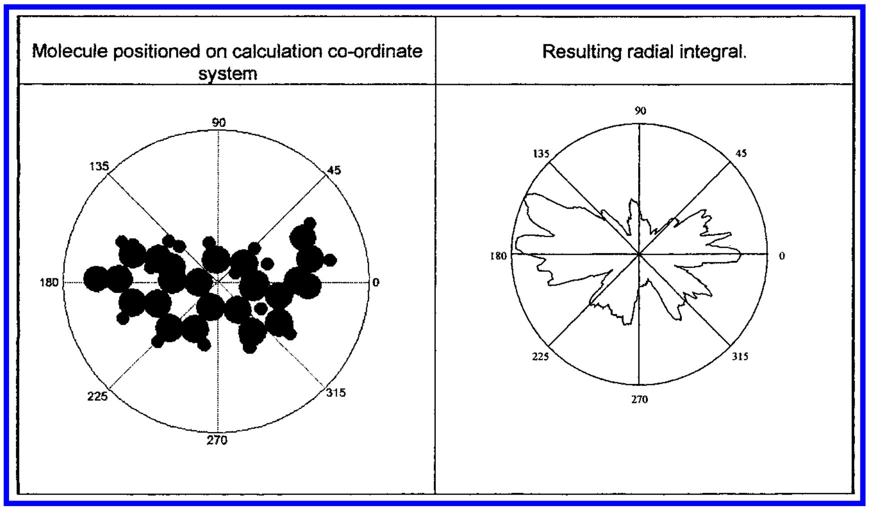

Molecule positioned on calculation co-ordinate system
Resulting radial integral.
90
45
0
315
270
225
180
135
90
45
0
315
270
225
180
135

Figure 3. Example of radial integration when applied to a steroid molecule.

Chart 1

$$\eta_0 = \mu_{20} + \mu_{02}$$

$$\eta_1 = \sqrt{|(\mu_{20} - \mu_{02})^2 + 4\mu_{11}{}^2|}$$

$$\eta_2 = \sqrt{|(\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2|}$$

$$\eta_3 = \sqrt{|(\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2|}$$

$$\eta_4 = \sqrt[4]{\left|\begin{matrix}(\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] + \\ (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03})[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2]\end{matrix}\right|}$$

$$\eta_5 = \sqrt[3]{(\mu_{20} - \mu_{02})[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] + 4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03})}$$

$$\eta_6 = \sqrt[4]{\left|\begin{matrix}(3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] + \\ (3\mu_{12} - \mu_{30})(\mu_{21} + \mu_{03})[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2]\end{matrix}\right|}$$

the first two invariant central moments to implement a crude character recognition system which worked remarkably well.

Utilizing the invariant central moments is fairly straightforward. All we have to do is to project our molecular property onto a grid and run through the calculations detailed above. This gives us a seven-dimensional vector $\eta_{A,2D}$ which represents the distribution of $\rho_{A,2D}(x,y)$. The molecule to be compared can also be subjected to the same treatment, remembering that we no longer have to bother aligning A and B to get the correct answer. This yields a second seven element vector $\eta_{B,2D}$. The similarity, or more properly in this case the distance between the two molecules, is then given by the Euclidean distance of these two vectors in seven-dimensional space:

$$D_{A,B} = \frac{|\eta_{A,2D} - \eta_{B,2D}|}{\sqrt{|\eta_{A,2D}||\eta_{B,2D}|}}$$

The above equation can be calculated in a fraction of the time required for the traditional Carbo or Hodgkin indices.

RADIAL INTEGRATION

This powerful technique involves projecting our molecular property $\rho(x,y)$ upon a calculation coordinate system. As we have seen before we need not worry about where we perform this projection as any arbitrary offset is easily accounted for by evaluating $\bar{x}$ and $\bar{y}$.

Now that we have centered the property on the calculation coordinate system let us proceed to integrate $\rho(x,y)$ along radii moving from the origin to a distance at which the property is deemed to be zero. We then store this value in an array at the offset corresponding to that angle, as illustrated in Figure 3.

The radial integral can be expressed as

$$R_A(\theta) = \int_0^R \rho_{A,2D}(r,\theta)\mathrm{d}r$$

To compare two molecules all we have to do is compare their radial integrals. However, this is not as simple as perhaps it initially seems for we still have done nothing about the relative orientations of the two molecules. To get around this problem we utilize a branch of mathematics known as correlation theory. Correlation theory is based on an equation known as the correlation integral z_{AB}, which in our case can be stated as

$$z_{\mathrm{AB}}(\Delta\theta) = \int R_{\mathrm{A}}(\theta) R_{\mathrm{B}}(\theta + \Delta\theta)\mathrm{d}\theta$$

$\Delta\theta$ is being used to rotate the radial integral of molecule B around so that comes into and out of alignment with molecule A. When the molecules are perfectly aligned z_{AB} will have its maximum value; therefore, all that we have to do is evaluate the correlation integral for all values of $\Delta\theta$ and search for the maximum value. Assuming that we desire an accuracy of 360/N degrees for our search this process would appear to be of the order N^2. However the fundamental reason we are carrying out the nonlinear mapping is to gain speed in searching through a large database so it makes some sense to see if we can perform the search faster. It transpires that the method for carrying out the fast search is very simple provided that we make N an integer power of 2.

Consider forming the Fourier transform[8] of the correlation integral, which we shall denote as $\mathbf{Z_{AB}}(f)$. This yields

$$\mathbf{Z_{AB}}(f) = \int z_{\mathrm{AB}}(\Delta\theta) e^{-i2\pi f\Delta\theta}\mathrm{d}\Delta\theta = \int [\int R_{\mathrm{A}}(\theta) R_{\mathrm{B}}(\theta + \Delta\theta)\mathrm{d}\theta]\mathrm{e}^{-i2\pi f\Delta\theta}\mathrm{d}\Delta\theta$$

Now let us assume that the order of integration can be changed in the rightmost expression

$$\mathbf{Z_{AB}}(f) = \int R_{\mathrm{A}}(\theta)[\int R_{\mathrm{B}}(\theta + \Delta\theta)\mathrm{e}^{-i2\pi f\Delta\theta}\mathrm{d}\Delta\theta]\mathrm{d}\theta$$

Writing $\sigma = \theta + \Delta\theta$ we get

$$\mathbf{Z_{AB}}(f) = \int R_{\mathrm{A}}(\theta)[\int R_{\mathrm{B}}(\sigma)\mathrm{e}^{-i2\pi f(\sigma-\theta)}\mathrm{d}\sigma]\mathrm{d}\theta$$

$$\mathbf{Z_{AB}}(f) = \int R_{\mathrm{A}}(\theta)\mathrm{e}^{i2\pi f\theta}[\int R_{\mathrm{B}}(\sigma)\mathrm{e}^{-i2\pi f\sigma}\mathrm{d}\sigma]\mathrm{d}\theta$$

The term in brackets can be seen to be nothing more than the Fourier transform of the radial integral of molecule B thus

$$\mathbf{Z_{AB}}(f) = \mathbf{R_B}(f)\int R_{\mathrm{A}}(\theta)\mathrm{e}^{i2\pi f\theta}\mathrm{d}\theta$$

The remaining integral term is simply the complex conjugate of the Fourier transform of the radial integral of molecule A, $\mathbf{R_A}(f)$. Hence we have established the important result that

$$\mathbf{Z_{AB}}(f) = \mathbf{R_A^*}(f)\mathbf{R_B}(f)$$

This implies that the complete correlation function can be obtained from the inverse Fourier transform of $\mathbf{Z_{AB}}(f)$.

Initially all of this information seems somewhat underwhelming as the presence of two Fourier transforms and one inverse Fourier transform must surely increase the complexity of the algorithm. In fact this is not so. In the mid 1960s Cooley and Tukey,[9] showed that a fast Fourier transform of N data points could be carried out in a process which was of the order $N\log_2(N)$, provided that N was an integer power of 2.

Our process contains two Fourier transforms to calculate $\mathbf{R_A}$ and $\mathbf{R_B}$, one multiplication loop to form $\mathbf{Z_{AB}}$, an inverse Fourier transform to restore $\mathbf{z_{AB}}$, and a straightforward scan to find the maximum value, which for $N = 512$ represents approximately a 20-fold improvement in computational requirements over a straight evaluation of the correlation integral for all $\Delta\theta$.

Hence if we denote the similarity calculated from radial integrals as $S_{\mathrm{AB}}^{\mathrm{RI}}$ we can write the following equation

$$S_{\mathrm{AB}}^{\mathrm{RI}} = \frac{\max(z_{\mathrm{AB}}(\Delta\theta))}{\sqrt{\int R_{\mathrm{A}}^2(\theta)\mathrm{d}\theta \int R_{\mathrm{B}}^2(\theta)\mathrm{d}\theta}}$$

The denominator, once more, just scales the result to lie between 0 and 1. It adds only complexity 2N to the process and is therefore insignificant. Furthermore there are some indications that normalization is not all that important and could be omitted.

RADIAL SCANNING

This final technique is extremely similar to that of radial integration. Consider, again, projecting the molecular property onto a calculation surface and adjusting the center of the calculation system to the center of the property distribution as previously described. We now draw a circle around the property distribution, whose radius is such that the property can be deemed to be negligible. We then proceed to work in along a certain radius until the properties value exceeds a certain threshold. We then store this distance from the origin into the corresponding angle of the radial scan array as in Figure 4.

The resulting radial scan arrays of two molecules can then be compared in an identical manner to that used for the radial integral data, which provides it with the required rotational independence.

SPEED CONSIDERATIONS

All three techniques offer translation and rotational invariance for the two molecules being compared.

Whilst the calculation of the invariant central moments, radial integral, and radial scan arrays from the property of the molecule $\rho_{\mathrm{2D}}(x,y)$ may take a considerable time to complete, the results of these calculations are sufficiently small to be stored as part of a molecular database and will be available for immediate utilization by a searching algorithm.

For example, there are only seven invariant moments. Assuming a real number requires eight bytes of storage this translates into only 56 bytes of data being required to describe the complete molecule. By a similar argument we can show that the radial integral and radial scans would both require around 4k Bytes of storage each. These numbers should be compared with a three-dimensional Carbo index. If we assume a molecular volume of 10 cubic Å with 0.1 Å resolution along each axis we can see that we would require 8M bytes of storage to be allocated in our database for every molecule. Given that a database may be required to hold many thousands or tens of thousands of compounds, we see that the storage requirements begin to run into hundreds of GBytes.

APPLICATIONS

The main advantage of the three techniques for molecular comparison outlined above is their invariance to both translation and rotation. To investigate the level of this invariance the two-dimensional representation of steroid 1,

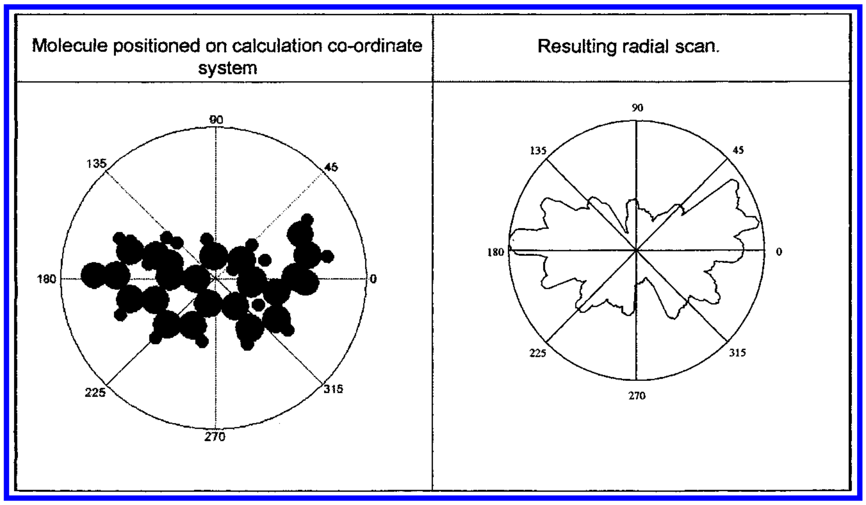


Figure 4. Example of radial scanning when applied to a steroid molecule.

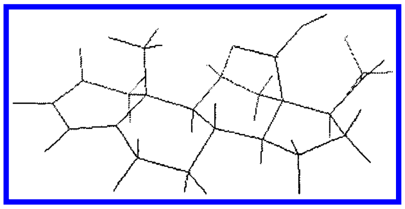


Figure 5. 3D representation of steroid 1.

whose structure is shown in Figure 5, was compared with itself, both before and after translation and rotation.

The results are summarized in Table 1. As can be seen from the table the three methods of similarity calculation show exceptional invariance to both translation and rotation. The slight discrepancies which occur on the rotated data can be put down to rounding errors. The 2D representation had to be projected onto a grid whose vertices were, for addressing purposes, at integer positions. Once the structure was rotated, those parts of atoms which did not fall precisely on a grid square were lost. As can be seen the effect is negligible.

APPLICATIONS TO THE STEROID DATA SET

The benchmark set of steroids is given in Table 2 and has been the basis of several studies.[10,11]

The three-dimensional structures of the above steroids were first converted into two-dimensional representations using the NLM. The resulting structures were then used to calculate the similarity information. The calculation of the three, 31 × 31 similarity matrices, took 2 min 30 s on a PC486-33.

TESTING THE SELF-CONSISTENCY

One of the most important tests of our two-dimensional similarity measurement techniques is whether the results from the three different methodologies are consistent with each other since they are supposed to be different measurements of the same thing.

We present the calculated similarity matrices in Figure 6. The brighter the square, the more similar the two molecules (note that this implies that the invariant moment matrix has been negated).

From these images we can see immediately that all three methods of measuring molecular similarity give largely the same results. Notable points are as follows:

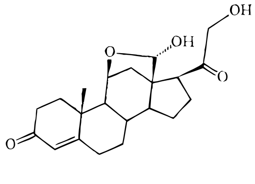
1. The radial integral and radial scan similarity matrices are visually more alike than the invariant moment similarity matrix.

2. The invariant moment similarity matrix has significantly lower "contrast" than the radial integral and radial scan matrices implying that as a technique, invariant moment calculations are less sensitive than radial integration and radial scanning.

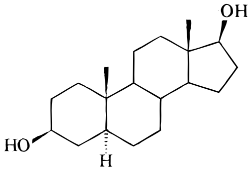
Table 1. Results of Translation and Rotational Invariance Testing

No Operation								
Molecular distance matrix from invariant moment calculation			Molecular similarity matrix from radial integral calculation			Molecular similarity matrix from radial scan calculation		
	steroid 1	steroid 1		steroid 1	steroid 1		steroid 1	steroid 1
steroid 1	0	0	steroid 1	1	1	steroid 1	1	1
steroid 1	0	0	steroid 1	1	1	steroid 1	1	1
Translation by 10 Å in the *x* and *y* Direction								
	steroid 1	steroid 1(T)		steroid 1	steroid 1 (T)		steroid 1	steroid 1 (T)
steroid 1	0	0	steroid 1	1	1	steroid 1	1	1
steroid 1 (T)	0	0	steroid 1 (T)	1	1	steroid 1 (T)	1	1
Anticlockwise Rotation of 45°								
	steroid 1	steroid 1 (R)		steroid 1	steroid 1 (R)		steroid 1	steroid 1 (R)
steroid 1	0	0.01008	steroid 1	1	0.997	steroid 1	1	0.9944
steroid 1 (R)	0.0100806	0	steroid 1 (R)	0.99702	1	steroid 1 (R)	0.994369	1

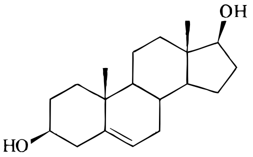
Table 2. Steroid Structure Data Set Used in the Calculation of the Similarity Matrices



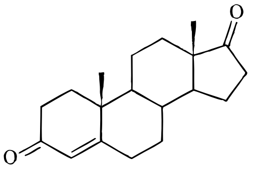
1



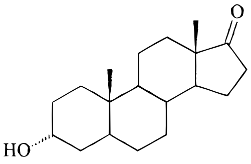
2



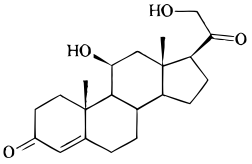
3



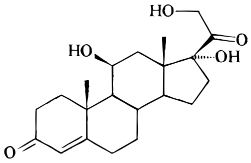
4



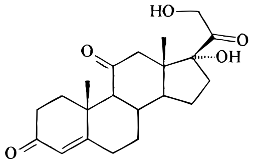
5



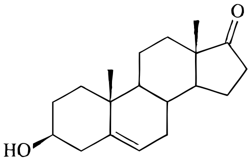
6



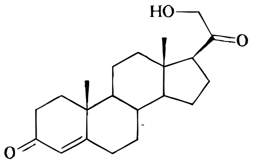
7



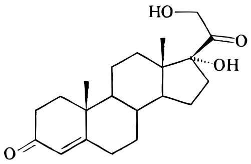
8



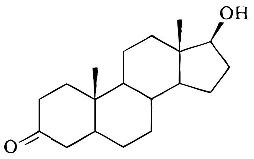
9



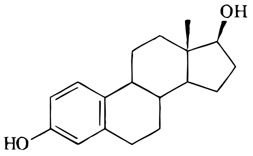
10



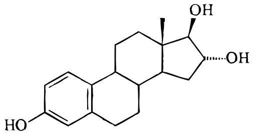
11



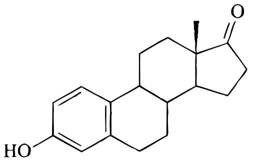
12



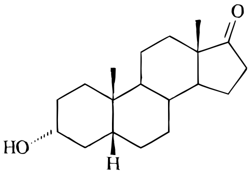
13



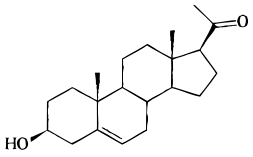
14



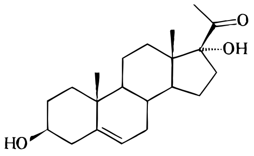
15



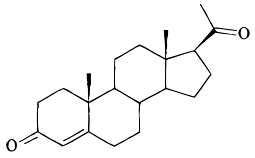
16



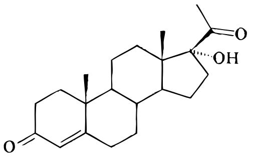
17



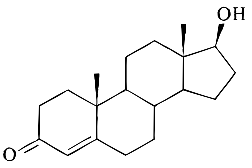
18



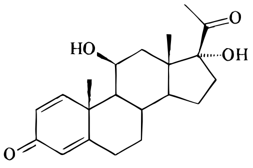
19



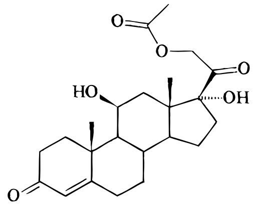
20



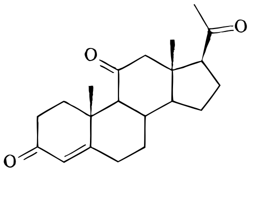
21



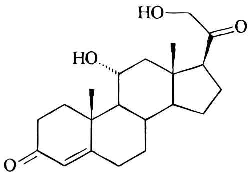
22



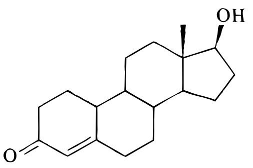
23



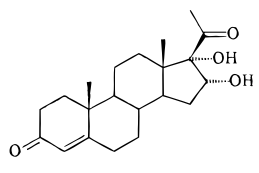
24



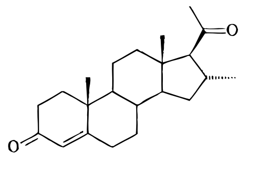
25



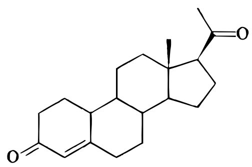
26



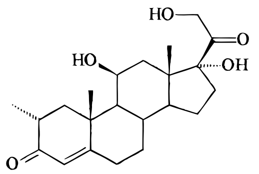
27



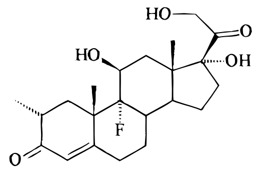
28



29



30



31

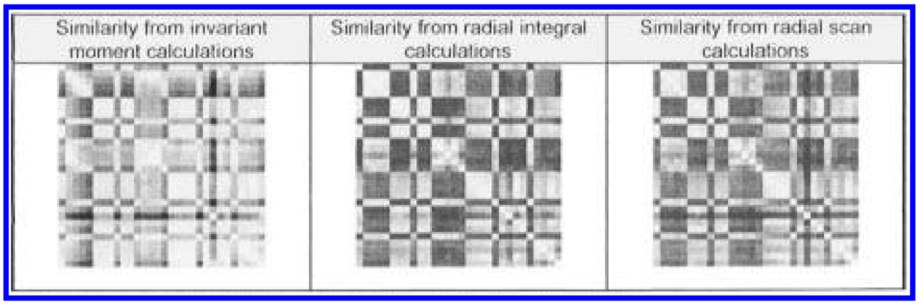


Figure 6. Images of the three 31 × 31 shape similarity matrices generated by the various 2D similarity measures introduced.

Table 3. Results from the Correlation of the 2D Techniques

	similarity from invariant moment calculations	similarity from radial integral calculations	similarity from radial scan calculations
similarity from invariant moment calculations	1	0.759	0.825
similarity from radial integral calculations	0.759	1	0.929
similarity from radial scan calculations	0.825	0.929	1

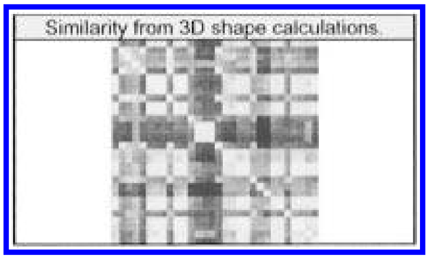


Figure 7. 3D shape similarity matrix calculated by ASP.

Table 4. Correlation of 2D and 3D Shape Similarity Matrices

	2D invariant moment calculations	2D radial integral calculations	2D radial scan calculations
3D shape calculations	0.622	0.731	0.735

The numerical correlation between the similarity matrices is given by Pearson's equation

$$\chi_{\mathrm{Data1,Data2}} = \frac{\sum\sum(\mathrm{Data1}-\mathrm{mean(Data1)})(\mathrm{Data2}-\mathrm{mean(Data2)})}{\sqrt{\sum\sum(\mathrm{Data1}-\mathrm{mean(Data1)})^2\sum\sum(\mathrm{Data2}-\mathrm{mean(Data2)})^2}}$$

the double summation being a consequence of correlating both the rows and the columns of the similarity matrices. The results of this correlation are shown in Table 3 where the high values for the correlation coefficient demonstrate the extremely high level of agreement between the various 2D similarity calculations.

COMPARISON WITH 3D SIMILARITY CALCULATIONS

We should expect a high degree of correlation between the 2D and 3D shape similarity metrics if they contain similar information. To test this the 31 steroids shown were passed into ASP,[12] and a calculation of the molecular similarity was made on the basis of three-dimensional shape. The results obtained are shown in Figure 7.

Visually the 3D similarity matrix appears to contain the same, gross structure as the 2D matrices. In order to gain a more detailed impression of how the 2D and 3D techniques compared, an analysis of the correlation was performed on a per molecule basis, as shown in Figure 8 as well as from the complete similarity matrix, Table 4. Figure 8 shows how each row in the 2D similarity matrix correlates with the corresponding row in the 3D similarity matrix.

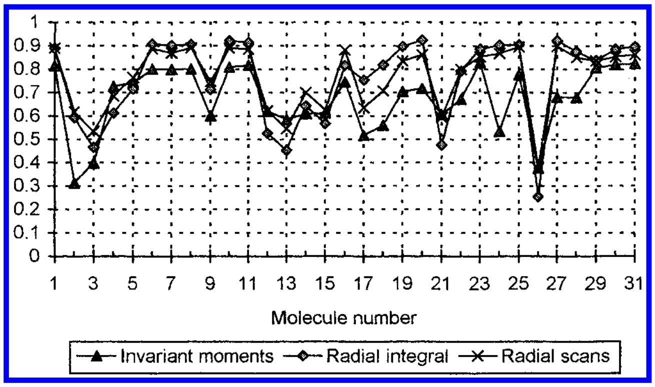


Figure 8. Molecule by molecule correlation of 2D and 3D shape similarity matrices.

There are several important points to note about Figure 8 and the overall correlation coefficients in Table 4.

1. The overall correlation coefficients between the 2D and 3D measurements are only about 70%. To investigate whether this was a significant feature a second 3D similarity calculation was performed using ASP, with electrostatic potential (ESP) being the similarity metric. The resulting similarity matrix was then compared with the 3D shape similarity matrix. According to Pearson's equation these two matrices only had a 70% correlation. As the discrepancy between the 2D and 3D calculations is no larger than that between two established 3D techniques there does not seem anything too alarming in the above figures.

2. The majority of the 2D and 3D shape similarity results agree with a correlation which is better than 80%. However the overall correlation between the 2D and 3D matrices is lowered by a few "rogue molecules" which have significantly lower correlations (for example molecule 26). It was initially felt that this might be due to the default choice in the center of interest for the three similarity techniques, for it is notable that the molecules with lower correlations are those with no side chain attached to the five-membered ring of the steroid core. However after studying the default choice for the center of interest of all 31 steroids there appeared to be no particular reason to believe that this was the case. Indeed all three techniques showed excellent invariance to the very slight movement of the center of interest caused by the side chains.

Instead it is felt that the lower correlation obtained for certain molecules is a fundamental property of the algorithms used. We have explained previously that most 3D shape similarity algorithms simply count up the positions where both molecules are present at the same time; however, they do not penalize the molecules' similarity where they differ. This contrasts with the techniques shown above where differences between the molecules causes a reduction in the similarity of the two molecules.

Going back to the steroid molecules, we can see that the 2D and 3D measurements will agree most strongly when comparing two molecules with side chains or two molecules without side chains. However for the cross comparison, one molecule with a side chain and one without, the agreement will not be so strong. In our series of steroids we have 19 molecules with side chains and 12 molecules without, hence we can see that we should expect stronger correlation between the 2D and 3D techniques for those molecules with side chains from the five-membered ring as shown above.

It is, of course, understood that it is not a measurement of similarity which is important, rather how a set of descriptors can be used to analyze and predict a molecule's activity. The utilization of molecular similarity matrices for such analyses is a very well documented aspect of the field. We have carried out several such investigations on a variety of molecule sets. Thus far the results of these tests have been extremely promising, with errors in the predicted activity being on a par with comparable 3D-QSAR studies.

SUMMARY

Invariance to Displacement in Relative Configuration Space. This was our prime consideration in the development of these two-dimensional similarity metrics. Firstly, we have shown that the technique can be made utterly invariant to translation.

Secondly, we have shown that the invariance to rotation can achieve an accuracy which is limited by both the fineness of the 2D grid we project the property onto (i.e., to the amount of memory we are prepared to use in the calculation and storage of the 2D representation data) and ultimately to the precision of the machine's floating point unit. These effects were shown to be negligible on a standard PC and are not felt to pose a serious problem on other computational platforms.

Speed of Comparison. Once the similarity data for each molecule has been calculated, an operation that only has to be performed once in the lifetime of any molecular structure, the comparison can be carried out extremely quickly. In the case of the formation of three $31 \times 31(=961)$ element similarity matrices the time taken was a small fraction of that required for the full 3D calculations.

Consistency of 2D Results. Our three techniques for the evaluation of 2D similarity have been shown to be entirely consistent with each other. This being especially true of the radial integration and radial scanning methods.

Consistency of Results between 2D and 3D Techniques. We have shown that the agreement between the 2D and 3D techniques is at least as good as the agreement between two commonly used 3D similarity metrics. Secondly, we have shown that any differences between the techniques are due to fundamental differences in philosophy rather than an obvious limitation of the 2D comparison algorithm.

CONCLUSIONS

Current 3D similarity techniques consume so much computing power that attempting calculations on large data sets is impossible. In order to avoid coming to a halt it is therefore necessary to radically alter the methods of performing the calculations. As promised in the previous paper in this issue we have shown how accurate 2D representations of molecules provide such a technique.

There is, of course, much work to be done. Firstly, we have only scratched the surface of available pattern recognition algorithms. Secondly, investigation is required into how different conformations of flexible molecules can affect the results of 2D calculations. Our currently favored technique for tackling flexible molecules is to consider each low energy conformer as a distinct structure. The speed and storage advantages bestowed upon us by the 2D similarity techniques assist greatly with the increased number of effective compounds. Finally we must consider how best to analyze the huge tables of results generated by large scale similarity calculations, where questions of molecular diversity arise.

ACKNOWLEDGMENT

D.D.R. would like to thank David Robinson for his helpful discussions during the development of the above techniques and for his useful comments during the preparation of this article.

REFERENCES AND NOTES

(1) Carbo, R.; Leyda L.; Arnau, M. An Electron Density Measure of the Similarity between two compounds. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189.

(2) Hodgkin, E. E.; Richards, W. G. Molecular Similarity based on Electrostatic Potential and Electric Field. *Int. J. Quantum Chem. Quantum Biol. Symp.* **1987**, *14*, 105−100.

(3) Meyer, A. M.; Richards W. G. Similarity of Molecular Shape. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 426−439.

(4) Pratt, W. K. Digital Image Processing, 2nd ed.; Wiley Interscience.

(5) Gonzales, R. C.; Woods, R. E. Digital Image Processing; Addison Wesley.

(6) Hu, M. K. Visual Pattern Recognition by Moment Invariants. *I.R.E. Trans. Inf. Theory.* **1962**, 179−187. (See also above two references).

(7) Papoulis, A. *Probability, Random Variables and Stochastic Processes*; McGraw-Hill: New York, 1965.

(8) Brigham, E. O. *The fast fourier transform*; Prentice-Hall Inc.: 1974.

(9) Cooley, J. W.; Tukey, J. W. An algorithm for machine calculation of complex Fourier series. *Math. Comput.* **1956**, *19*, 297−301.

(10) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(11) Good, A. C.; So, Sung-Sau; Richards, W. G. Structure Activity Relationships from Molecular Similarity Matrices. *J. Med. Chem.* **1993**, *36*, 433−438.

(12) ASP (Automated Similarity Package); Oxford Molecular Limited: The Magdalen Centre, Oxford Science Park, Sandford on Thames, Oxford, OX4 4GA, United Kingdom.

CI970423T