Alignment of 3D-Structures by the Method of 2D-Projections

Daniel D. Robinson, Paul D. Lyne, and W. Graham Richards*

Physical and Theoretical Chemistry Laboratory, University of Oxford, South Parks Road, Oxford, OX1 3QZ, United Kingdom

Received October 24, 1998

The three-dimensional shape of a molecule can be analyzed by considering a series of two-dimensional projections. By comparing projections of two molecules a mechanism for the alignment of three-dimensional structures is derived. The efficiency of this two-dimensional comparison permits a sequential search over alignment space thereby eliminating misconvergence problems that plague many existing automated alignment procedures. Examples of molecular alignment by this method are illustrated together with several variations on the basic procedure. These variations permit a wide variety of structures to be considered. The existence of a reliable and efficient automated alignment methodology has important ramifications for QSAR studies and general chemical informatics.

INTRODUCTION

Molecular alignment is of crucial importance in the fields of quantitative structure−activity relationship (QSAR) generation, molecular diversity, and structural database searching.[1,2] Currently there are several methods available for molecular alignment. Techniques used to perform rigid alignments of molecules include the atom based methods that rely on a correspondence of molecules to predefined atomic positions or pharmacaphores and have only limited use.[3] More sophisticated approaches use electrostatics, shape, or molecular similarity indices to produce an optimal superposition of molecules.[4−6] These attempt to optimize the similarity of the two structures by a series of rigid translations and rotations. In general such methodologies are more robust than the atom based techniques. Unfortunately they are also more complex to implement successfully as they require an extremely robust optimization algorithm if the system is not to get caught on local extrema of the similarity hypersurface. One of the more reliable optimization algorithms has been implemented by Parretti et al.[6] who used a Monte Carlo based optimizer for the alignment procedure, rather than the downhill-only simplex method more normally employed. However despite a considerable amount of effort the Monte Carlo algorithm can still get trapped in local extrema. In addition the Monte Carlo algorithm is not very time efficient. Other approaches use approximate alignments to facilitate fast database searching,[7] while the issue of molecular flexibility has been taken into consideration in a limited number of cases.[8−10]

The alignment of objects has received considerable attention in the fields of machine vision and machine intelligence, where inefficiencies in searching algorithms are not tolerable. Wallace and Mitchell[11] provide an example of three-dimensional object alignment. In their case they wished to track the movement of a three-dimensional object, specifically an aircraft, as it maneuvered. To do this they generated a library of images of the aircraft from different known positions and encoded these images in such a manner as to enable facile comparison between this library and a real-life scene. When an aircraft was observed, its image was compared to every image in the library. Through this method they could tell not only what type of aircraft they were looking at but also its position in space relative to them. Clearly once an object's alignment relative to a model is known realigning that object with the model is easy. The existence of such algorithms provided the starting point for the investigation of the use of 2D-projections for the fast alignment of molecules.

THEORY

The analysis of three-dimensional shape can be roughly classified into two categories, local shape-analysis, and global shape-analysis. Local shape-analysis is by far the more complex of the two. It involves us being able to align and compare two structures when only a small fraction of one structure is present in the other. Various statistical and syntactic hybrid methodologies have attempted to solve the local shape-analysis problem, and we will report on our investigations using these methodologies in a subsequent publication.

This paper focuses on the simpler category of global shape-analysis, where all of one structure is to be matched with the whole of another structure. This is done in order to evaluate whether such radically "nonchemistry" based algorithms have any promise in computational chemistry. In addition, this class of analysis closely overlaps the currently available alignment techniques described above, which almost invariably attempt to align one complete molecular structure with another.

Alignment of two rigid three-dimensional structures requires the optimization of six degrees of freedom; three translational and three rotational. For global shape-analysis, as a simplification it can be assumed that the mean position of each structure to be aligned is at a comparable position in each structure. This assumption is clearly only valid over a limited range of structures, such as those found in a series of similar molecules. While this limitation may appear quite restrictive, it is found in practice not to be a problem. Most

10.1021/ci9803379 CCC: $18.00
Published on Web 04/22/1999

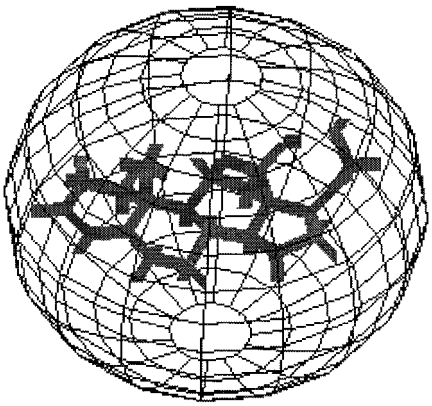


Figure 1. Illustration of a molecule centered within its sphere.

series of molecules under consideration are sufficiently similar for the approximation of comparable means to hold. Those pairs of structures that are sufficiently different to cause problems are better handled by local shape-analysis techniques. Given that the mean positions of each structure are comparable the three translational degrees of freedom can be eliminated, thereby reducing the complexity of the problem.

The remaining three rotational degrees of freedom are dealt with in an elegant manner as follows. Consider an imaginary wire sphere around each molecule, whose radius is such that all structures to be considered fit easily within it (Figure 1). Excluding the very polar regions of the sphere the lines of longitude and latitude form a rectangular matrix, which although not evenly spaced, can be handled easily on a computer. At every position where the lines of longitude and latitude intersect a 2D picture of the molecule when viewed along the axis connecting the center of the sphere to the intersection point is recorded. This axis is the viewing axis. The picture is that of the molecular shape generated by rendering each of the atoms of the molecule as a sphere. This picture is stored at the position specified by the lines of longitude and latitude in a library of images of the molecule. A low-resolution example of such a library is provided in Figure 2. The lines of longitude and latitude hence specify two of the three remaining degrees of rotational freedom. The final degree manifests itself as a rotation of the picture stored at each position of the library, i.e., a rotation about the viewing axis.

The task of alignment therefore becomes a matter of comparing each image generated by one molecule with one or more images from a second molecule, in a manner that is rotationally independent. The pair of images that exhibit the minimum difference, or the maximum similarity, specify the transformation that aligns the two structures. Techniques for comparing images in a rotationally invariant manner are widely known and documented in the field of image processing.[12,13] The applications of these techniques to molecular alignment are outlined here and described in detail elsewhere.[14,15]

ROTATIONALLY INVARIANT IMAGE COMPARISON

To compare two images in a rotationally independent manner a measure of the distance between the two images is required. This distance metric must not be affected by the fact that one of the images could be rotated with respect to the other. Two such metrics are provided by the related techniques of radial integration and radial scanning.

Figure 2 depicting a complete library of 2D-projection images is a representative example of the data that needs to be analyzed. The first step in the comparison of an image with another image of a different molecule is to find the center of the two-dimensional projection of the molecule on each image. Either the center can be taken as being the center of the image on which the projection is stored or alternatively the center can be evaluated by the method of moments.[16] This is the technique currently implemented in the software. Equation 1 shows the form of the moment generating equation. The value $f(x,y)$ represents the data stored on the image where $f(x,y) = 0$ outside the molecule and $f(x,y) = 1$ inside the molecule. Equations 2 and 3 show how the center of the two-dimensional representation ($\bar{x}$, $\bar{y}$) can be found from the moment generating equation.

$$\mu_{p,q} = \sum_x \sum_y x^p y^q f(x,y) \tag{1}$$

$$\bar{x} = \frac{\mu_{1,0}}{\mu_{0,0}} \tag{2}$$

$$\bar{y} = \frac{\mu_{0,1}}{\mu_{0,0}} \tag{3}$$

Having found the center of the two-dimensional projection stored on the image it is necessary to encode the projection in a manner that is suitable for rapid comparison. This is done by starting from the center of the two-dimensional projection and working out toward the edge of the image along a given angle θ. While traveling along this line the function $f(x,y)$ is accumulated. On reaching the edge of the image the accumulated value is stored in an array RI at the position corresponding to θ. This is repeated across all values of θ until a complete circle has been completed. The array RI is the radial integral array. Clearly RI is periodic in θ such that $\mathrm{RI}(\theta) = \mathrm{RI}(\theta+2\pi)$. The radial integral array of one two-dimensional projection RI1 can be compared with that of another RI2 by calculating the absolute difference. Because the functions are periodic in θ eq 4 can be used to find the rotation angle ϕ which minimizes the absolute difference $\mathrm{Diff}(\phi)$ between RI1 and RI2.

$$\mathrm{Diff}(\phi) = \sum_\theta |\mathrm{RI}_1(\theta) - \mathrm{RI}_2(\theta + \phi)| \tag{4}$$

The value of $\mathrm{Diff}(\phi)$ is an indication of how well the two-dimensional projections compare. This is the distance metric, and the value of ϕ is the relative angle of the optimum alignment. ϕ represents the final rotational degree of freedom for the three-dimensional structures and completes all of the information we required to align the structures.

Radial scanning begins at the edge of the image and progresses toward the center along an angle θ. When the boundary of the two-dimensional projection is encountered, i.e., where $f(x,y)$ goes from 0 to 1, the distance of the boundary from the center is stored. Progressing through values of θ yields the radial scan array $\mathrm{RS}(\theta)$ which has identical properties to the radial integral array and can be utilized in an identical manner.

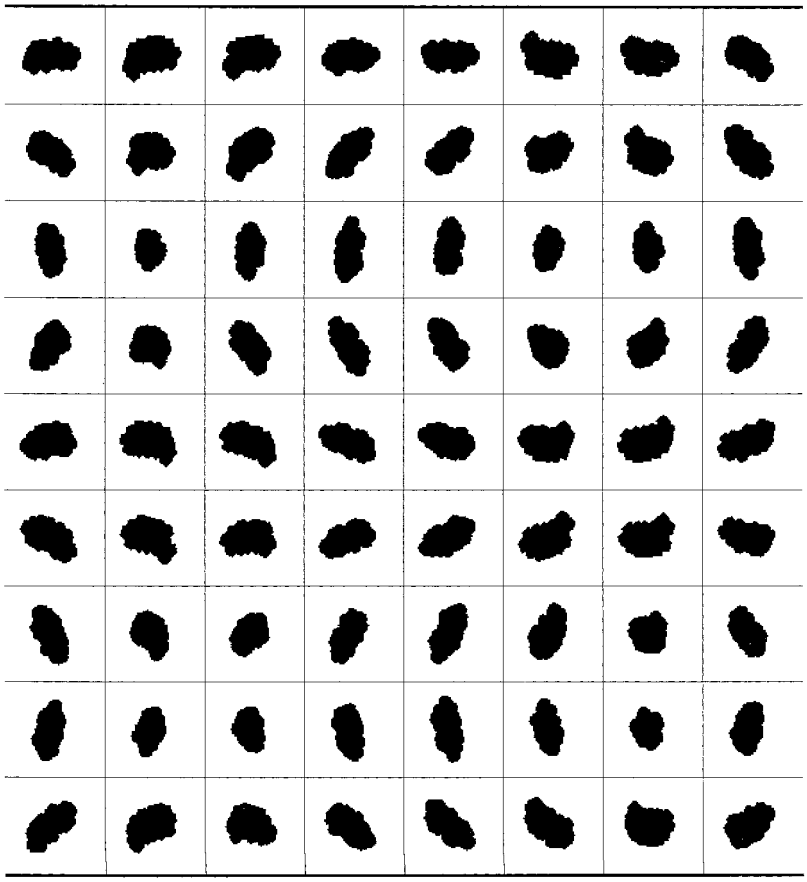


Figure 2. An example library of 2D-projections.

COMPUTATIONAL DETAILS

The method is implemented in Microsoft Visual Basic v5.0,[17] with the radial integrals and radial scans being written as COM objects within Microsoft Visual C++ v5.0 for speed.[18] The software is run under Microsoft Windows NT 4.0 on a Pentium II PC. Rendering of the molecule's shape is carried out using OpenGL 1.1, built into the Windows operating system.

RESULTS

Alignments can be performed in single image mode or multiple image mode and each of these is considered in turn. Every alignment is shown together with a "Fit" value. This quantity attempts to measure the quality of the alignment of the pair of molecules shown. The Fit value is calculated in several stages. A list is created containing the square distance of each atom in the larger molecule from the closest atom in the smaller molecule, the molecules size being judged from the number of atoms in the structure. The mean and standard deviation of these distances is then calculated. Those distances which are less than the mean distance plus two standard deviations are then used to calculate the RMS distance between the structures. This technique is used to attempt to give some meaning to an RMS distance when applied to molecules that do not possess an exact atom matching. Features that do not exist on one structure will have a higher distance than those features with an obvious pairing. Such features will be removed from the RMS calculation by the filter preventing them unfairly biasing the results.

Single Image Alignment. This mode involves mapping a single image of a molecule (the molecule to be aligned) to a complete library of images of another template molecule (the molecule we are aligning against). This is the fastest mode of operation for the software. To test this mode of operation sets of molecules with related structures were selected from a variety of well-known QSAR data sets. Their orientations were then randomized by multiplying the position vectors of the atoms of each molecule by a randomly selected three-dimensional rotation matrix. The software was then used to align the molecules within a given set.

Figure 3(a) shows the structure of the template molecule used in this first series of tests. Figure 4(c,d) shows the results of aligning this template molecule with other similar, randomly oriented structures. Visually the alignment is exceptionally good and can be used directly or as a starting point for a similarity optimizer, which would no longer face the problem of being trapped on local extrema. The quality of the alignment is backed up by the low value for the Fit parameter.

Multiple Image Alignment. While the single image based alignment is surprisingly reliable, a problem can arise when the view is not representative of the structure of the molecule to be aligned. For example if the view is "end on" very little of the structure of the molecule in the image is seen yielding an unstable alignment. To counteract this problem a complete library of images of one molecule can be compared with a complete library of images of a second molecule. To carry out this operation one image is taken from a library of 2D-projections of the template molecule and compared with a

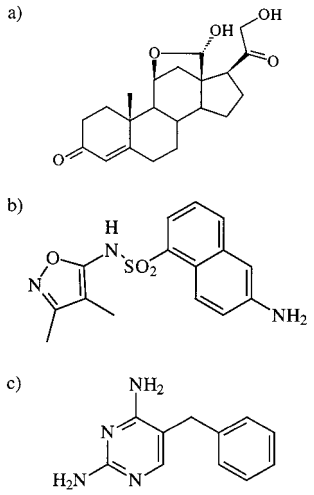


Figure 3. The template molecules used in the tests.

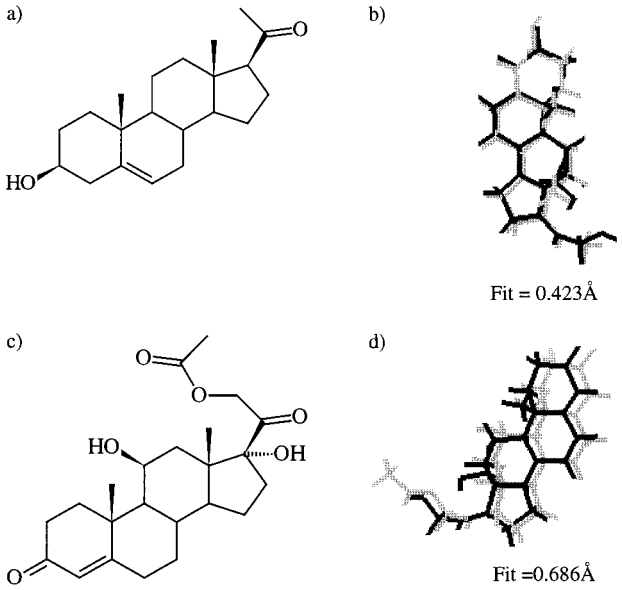


Figure 4. Results of the single image alignment.

complete library of images of the molecule to be aligned. This operation is repeated until all possible pairs of 2D-projections have been compared. The closest matching 2D-projections then specify the transformation that is used to align the structures. Obviously this is somewhat slower than the single image alignment; however, it has the advantage of being significantly more robust. In effect this type of alignment performs an iterative search across all of the rotational degrees of freedom in the system, since each image in a given library represents the molecule in one particular orientation.

As an example consider Figure 5(a,b), where the alignment of the molecule with the previous template has clearly failed when using a single image. Compare this with Figure 5(c) where the alignment based on multiple images is totally convincing both visually and in terms of the Fit value which is considerably smaller than that of Figure 5(b).

Variations. A number of modifications to the basic alignment methodology can be made. First a simple modification allows us to concentrate our efforts selectively on a promising potential alignment while skipping rapidly over regions where the molecules are obviously not aligned. The current implementation of the software performs this operation as follows. Initially a relatively coarse library is made, the resolution being set by the user, but typically a resolution of about 20° is used. When an area of misalignment is found the program continues the coarse comparison operation. However if a potentially good alignment is found the software generates some additional images at much smaller rotational increments and compares these in order to find the optimum alignment position more accurately.

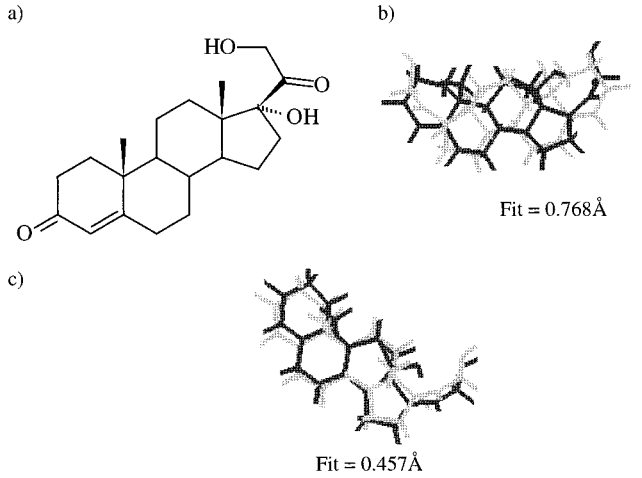
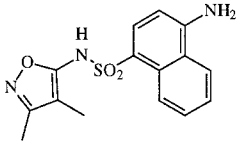


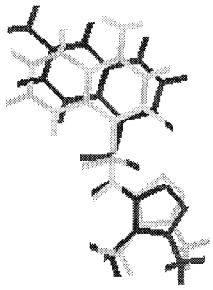
Figure 5. Illustration of the greater reliability of the multiple image based alignment.

Although far less frequent than the single image situation, the multiple image alignment can be fooled by views of molecules that are "end on". The reason for this is that the "end on" views have considerably less information to match with another "end on" view. Consequently it is much easier to find a match between these views relative to more informative views. We can circumvent such problems quite easily by favoring those images that contain more information about the complete molecular structure. We can discern such images by considering the variance of either the radial integrals or radial scans. In general the variance is much greater for high information images than for low information "end on" images, the only exception being planar or near planar molecules, which appear to be handled adequately without this option. By dividing the distance between the two images according to their variance we selectively favor matches between high information images relative to low information images.

Both of these techniques were applied in a second series of tests, whose template molecules are shown in Figure 3(b,c). An attempt was made to align the template structure in Figure 3(b) with the structures shown in Figure 6(a−f); Figure 3(c) was aligned with the structures in Figure 6(g−l). For both series of alignments an effort was made to increase gradually the structural diversity of the molecules being aligned. As can be seen this alignment methodology copes well with most changes in structure. For example the benzyl-sulfonamide structures shown in Figure 6(e,f) are still well aligned with the naphthyl-sulfonamide template. Indeed the only time the 2D-projection alignment technique can be seen to fail is in the case of Figure 6(l), which also exhibits a large value for the Fit parameter. This is not entirely surprising, the two structures being aligned are of considerably different size, in terms of counting rings Figure 6(l) is some 50% larger than the template. In this case our

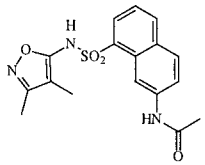
a)

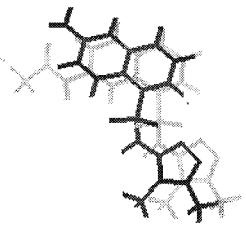




Fit = 0.489Å

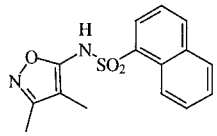
b)

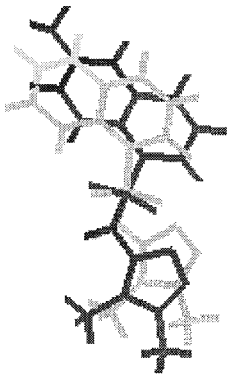




Fit = 0.963Å

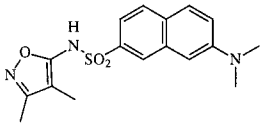
c)

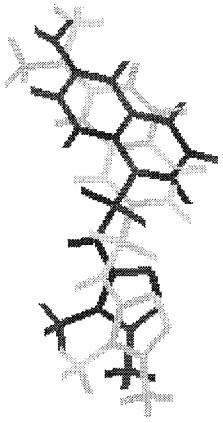




Fit = 0.664Å

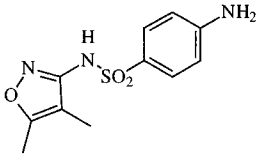
d)

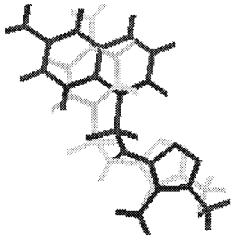




Fit = 0.843Å

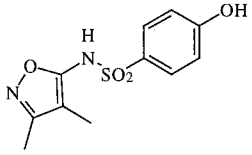
e)

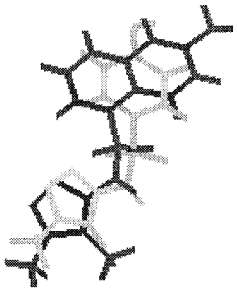




Fit = 0.845Å

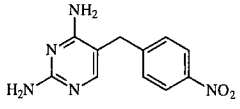
f)

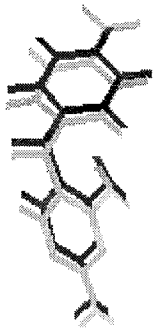




Fit = 0.814Å

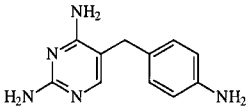
g)

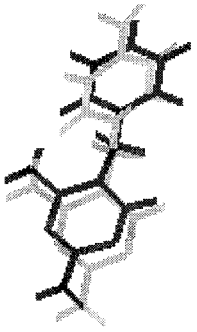




Fit = 0.429Å

h)





Fit = 0.515Å

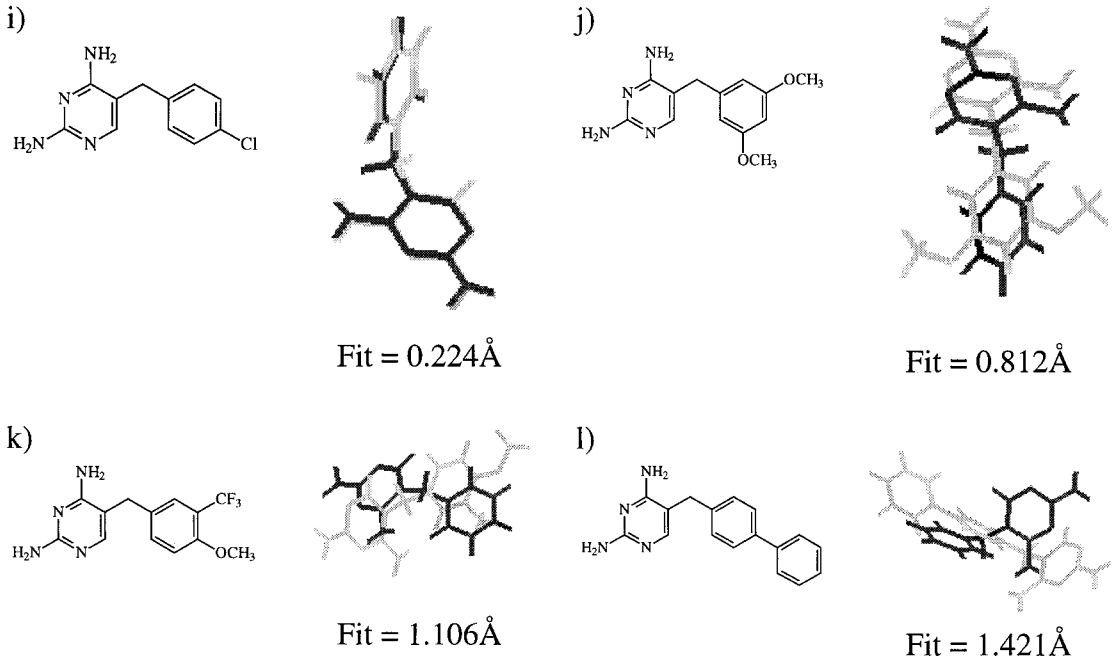


Figure 6. Results from the second set of tests.

assumption that the centers of each molecule are in a comparable position has broken down. We believe that such a disparity in size necessitates the utilization of an alignment methodology based upon local shape-analysis.

CONCLUSIONS

The techniques presented here have a number of features that are beneficial when dealing with the molecular alignment problem. First this technique can be used as the sole alignment procedure. As shown here, and proved in many other tests with different molecular test sets, alignment of similar three-dimensional structures via their two-dimensional projections is a reliable and efficient methodology. Alternatively this method of structure alignment can be used to help improve the efficiency of existing alignment algorithms. By using the output of this technique as the starting point for a similarity optimization alignment the majority of the local extrema which exist on the similarity hypersurface can be bypassed. This is expected to improve the reliability of such programs considerably. Furthermore this technique may assist with the fitting of flexible molecules; one could envisage screening a large number of potential conformations of a flexible molecule for their fit to a template. Only the most promising of these candidates would then be passed on to a flexible fitting routine, once more eliminating a large number of "misalignments" which would otherwise have to be considered.

In carrying out this study it is seen that existing tools based on chemical intuition are not necessarily the only tools that are capable of performing the task required, nor are the existing tools necessarily the best ones for the job, in terms of either their computational efficiency or their reliability. The field of pattern recognition contains many other novel and elegant methods, many of which have the potential to improve the techniques used for the complex three-dimensional world of chemical information handling and analysis. It is our intention to continue building upon the foundations of this work. We have already completed preliminary work on a local-shape analysis algorithm which is showing considerable promise, including docking structures into a defined site on a macromolecule. This should enable structures of greater diversity to be compared in a logical and consistent manner, a process that has important ramifications for the advancement of QSAR studies and general chemical informatics.

ACKNOWLEDGMENT

D.D.R. is supported by an EPSRC CASE studentship held in conjunction with Oxford Molecular Group PLC. This work was in part supported by the Wellcome Trust.

REFERENCES AND NOTES

(1) Klebe, G. Structural Alignment of Molecules. In *3D-QSAR in Drug Design. Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science Publishers: Leiden, The Netherlands, 1993; pp 173−199.

(2) Bures, M. G. Recent techniques and applications in pharmacophore mapping. In *Practical Applications of Computer Aided Drug Design*; Chariforn, P. S., Ed.; Marcel Dekker: New York, 1997; pp 39−72.

(3) Kim, K. H. Comparative Molecular Field Analysis (CoMFA). In *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Blackie Academic and Professional: Glasgow, 1995; pp 291−331.

(4) Kearsley, S. K.; Smith, G. M. An Alternative Method for the Alignment of Molecular Structures: Maximizing Electrostatic and Steric Overlap. *Tetrahedron Comput. Methodol*. **1990**, *3*, 615−633.

(5) Grant, J. A.; Gallardo, M. A.; Pickup, B. J. A Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape. *J. Comput. Chem*. **1996**, *17*, 1653−1666.

(6) Parretti, M. F.; Kroemer, R. T.; Rothman, J. H.; Richards, W. G. Alignment of Molecules by the Monte Carlo Optimization of Molecular Similarity Indices. *J. Comput. Chem*. **1997**, *18*, 1344−1353.

(7) Hahn, M. Three-dimensional shape-based searching of conformationally flexible compounds. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 80−86.

(8) Perkins, T. D. J.; Mills, J. E. J.; Dean, P. M. Molecular Surface-Volume and Property Matching to Superpose Flexible Dissimilar Molecules. *J. Comput. Aided. Mol. Des*. **1995**, *9*, 479−490.

(9) Jones, G.; Willet, P.; Glen, R. C. A Genetic Algorithm for Flexible Molecular Overlay and Pharmacophore Elucidation. *J. Comput. Aided. Mol. Des*. **1995**, *9*, 532−549.

(10) Lemmen, C.; Lengauer, T.; Klebe, G. FlexS: A Method for Fast Flexible Ligand Superposition. *J. Med. Chem*. **1998**, *41*, 4502−4520.
(11) Wallace, T. P.; Mitchell, O. R. Analysis of Three-Dimensional Movement Using Fourier Descriptors. *IEEE Trans. Pattern Anal. Machine Intell*. **1980**, *PAMI-2, no 6*.
(12) Pratt, W. K. *Digital Image Processing*, 2nd ed.; Wiley Interscience.
(13) Gonzales, R. C.; Woods, R. E. *Digital Image Processing*; Addison-Wesley.
(14) Robinson, D. D. Barlow, T. W. Richards, W. G. Reduced Dimensional Representations of Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 939−42.
(15) Robinson, D. D. Barlow, T. W. Richards, W. G. The Utilization of Reduced Dimensional Representations of Molecular Structure for Rapid Similarity Calculations. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 943−50.
(16) Hu, M. K. Visual Pattern Recognition by Moment Invariants. *I.R.E. Trans. Information Theory* **1962**, 179−187.
(17) Microsoft Corporation. One Microsoft Way, Redmond, WA 98052-5234.
(18) The software can be made available via ftp to interested parties.

CI9803379