

Reduced Dimensional Representations of Molecular Structure

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

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

Received March 14, 1997[®]

Two-dimensional representations of molecular structure may be generated from the three-dimensional coordinates by nonlinear mapping. These representations not only retain the look of familiar structural formulas but also incorporate the distance geometry information available in three-dimensional structures. The technique is shown to be trivial so long as the molecular structure is not essentially spherical and an objective test as to whether this is true is introduced, together with extensions for coping with such cases. The two-dimensional diagrams have great promise in handling molecular similarity and data searching with very large numbers of compounds since they permit the use of two-dimensional pattern recognition techniques.

INTRODUCTION

Molecular modelers are adept at recognizing representations of three-dimensional structures presented on graphics screens. Bench chemists prefer to think in terms of familiar structural formulas as are found in texts. In an earlier paper¹ we introduced the technique of nonlinear mapping² (NLM) as a way of presenting protein structures in two dimensions. That exercise showed that most structural features were retained in the two-dimensional picture including, to a surprising level of accuracy, the three-dimensional distance information.

Here we apply the same idea in a simpler and more rapid manner to small organic molecules. The value of this is firstly to provide a tool to make modeling programs available to bench chemists. Secondly the two-dimensional representations are capable of being compared to provide measures of molecular similarity³ far faster than is possible in three dimensions. Such accelerations are vital if quantitative similarity ideas are to be applied to the large numbers of compounds emerging from high throughput synthesis and combinatorial chemistry techniques.

THE THEORY OF THE NONLINEAR MAP

A molecule may be represented by a list of its atomic coordinates, measured relative to the molecular centroid. For a system of N atoms these may readily be stored in a $3 \times N$ matrix which we shall denote **P3**:

$$\mathbf{P3} = \begin{bmatrix} x_1 & x_2 & \dots & \dots & x_N \\ y_1 & y_2 & \dots & \dots & y_N \\ z_1 & z_2 & \dots & \dots & z_N \end{bmatrix}$$

We may use this list of atom positions to generate the distance matrix for the molecule in a process which will be familiar to anyone working in the field of protein structure analysis:

$$\mathbf{D3}_{ij} = \sum_{k=1}^3 (\mathbf{P3}_{k,i} - \mathbf{P3}_{k,j})^2$$

As can be seen our distance matrix simply contains the square of the Euclidean distance between the atoms. The distance matrix is an important concept as we can use it to regenerate the original 3D structure through distance geometry techniques.⁴

Let us now consider creating a new list of atomic coordinates in two dimensions which we shall denote **P2**, such that

$$\mathbf{P2} = \begin{bmatrix} x_1 & x_2 & \dots & \dots & x_N \\ y_1 & y_2 & \dots & \dots & y_N \end{bmatrix}$$

The initial values for the elements of **P2** may either be chosen randomly or, much better, by performing a principal component analysis on the initial 3D coordinates and discarding the data which lie along the tertiary eigenvector. We may use **P2** to generate a second distance matrix **D2** which is trivially defined as

$$\mathbf{D2}_{ij} = \sum_{k=1}^2 (\mathbf{P2}_{k,i} - \mathbf{P2}_{k,j})^2$$

Finally we may construct an error matrix **E** whose elements are given by

$$\mathbf{E}_{ij} = \mathbf{W}_{ij} (\mathbf{D3}_{ij} - \mathbf{D2}_{ij})^2$$

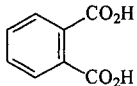
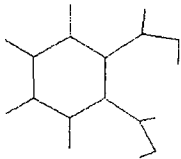
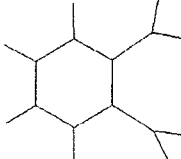
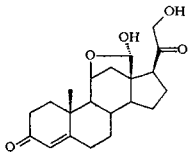

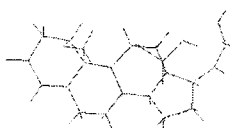
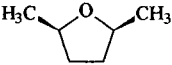

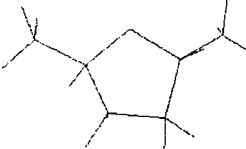
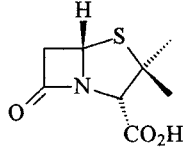
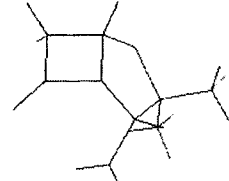
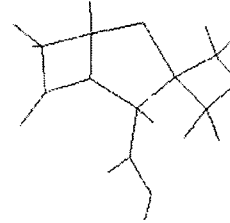
where **W** is a weighting matrix. This matrix did not appear in the original work and has been included to lend flexibility to the algorithm. This flexibility as we shall see later proves invaluable when dealing with molecules which are almost spherical in nature. We then set about an iterative procedure to minimize the total of the error matrix, by altering the positions of the atoms in the 2D representation **P2**. We have continued utilizing a “steepest descents” algorithm as in previous work. For this the following derivatives are required:

$$\frac{\partial \mathbf{E}}{\partial \mathbf{P2}_{ij}} = -4 \sum_{k=1}^N \mathbf{W}_{j,k} (\mathbf{P2}_{ij} - \mathbf{P2}_{i,k}) (\mathbf{D3}_{j,k} - \mathbf{D2}_{j,k})$$

$$\frac{\partial^2 \mathbf{E}}{\partial \mathbf{P2}_{ij}^2} = 4 \sum_{k=1}^N 2 \mathbf{W}_{j,k} (\mathbf{P2}_{ij} - \mathbf{P2}_{i,k})^2 - \mathbf{W}_{j,k} (\mathbf{D3}_{j,k} - \mathbf{D2}_{j,k})$$

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

Table 1. Numerical Results of Nonlinear Mapping for Selected Molecules

Structural formula	Three dimensional representation	NLM generated 2D representation.	Sphericity (Ω)	Error of 2D structure PCA, NLM (Å)
			0.16	0.12, 0.09
			0.24	0.31, 0.27
			0.38	0.28, 0.24
			0.43	0.35, 0.30

Each component of $\mathbf{P2}$ is then updated by applying the following equation

$$\mathbf{P2}_{i,j} = \mathbf{P2}_{i,j} - \eta \frac{\frac{\partial \mathbf{E}}{\partial \mathbf{P2}_{i,j}}}{\left| \frac{\partial^2 \mathbf{E}}{\partial \mathbf{P2}_{i,j}^2} \right|}$$

where η is a training factor which was kept at 0.4.

PROPERTIES AND EXAMPLES OF THE NONLINEAR MAPPING ALGORITHM

The NLM is one of many algorithms which is able to reduce the dimensionality of a set of data. Of this class of algorithm, perhaps the best known, and most widely used, is principal component analysis (PCA). It is helpful to consider the relative merits of both techniques in terms of their speed of execution and the overall error of the final 2D representation relative to the 3D structure. First of all let us consider the error in the final 2D representation, remembering that we are using the square of the Euclidean distance to calculate the distance matrices. We can see that a suitable definition for the error would be

$$\text{error} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \sqrt{(\sqrt{\mathbf{D3}_{i,j}} - \sqrt{\mathbf{D2}_{i,j}})^2}$$

Defining the error in this manner leaves it with the unit of length. In all cases which we have studied, the NLM is able to produce a 2D representation which is consistently 12–20% lower in error than a PCA analysis. Table 1 has

some results illustrating this point (it should be noted that all of these structures were generated with every element of the weighting matrix set to unity).

We believe that this improvement in error is important for two reasons. The first concerns the utilization of the 2D representations for quantitative molecular similarity calculations where the reliability of the results will be affected by the accuracy of the 2D representations in question.

The second point concerns the reversibility of the 3D to 2D process. It is obvious that the 2D structures require 33% less storage space than their 3D counterparts. In a small database of molecules this saving is insignificant; however, in a large database containing perhaps hundreds of thousands or maybe even millions of molecules, such a saving is to be taken more seriously, especially if one is to consider the transmission of such information across bandwidth limited communication devices such as telephone lines. However, there are, of course, times when only 3D coordinates will suffice. As we have stated distance geometry techniques can be used to reconstruct the 3D coordinates of a molecule from the distance matrix alone, calculable from the 2D coordinates. Clearly the regenerated structure can only be as good as the information which it is given. A 12–20% reduction in the error of the initial distance matrix will surely translate into an improved 3D structure.

The speed of the NLM is clearly a critical consideration, especially if we are to use it for large scale similarity measurements. Compared to PCA the NLM is slow, due largely to the recalculation of the matrix $\mathbf{D2}$ during each iteration cycle, a process which scales as $O(N^2)$. When started from a completely random set of 2D positions the

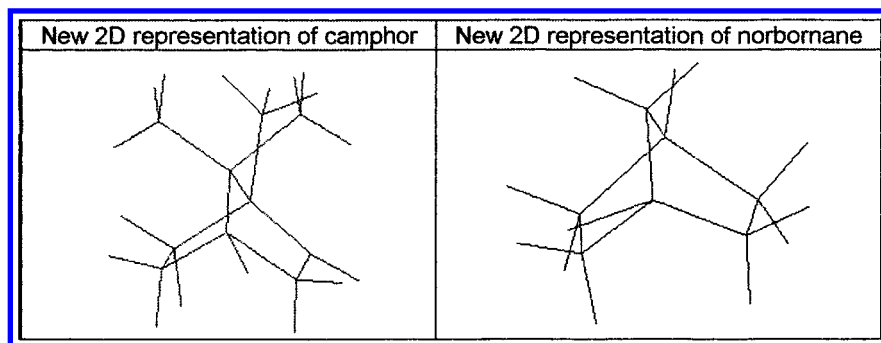
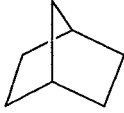
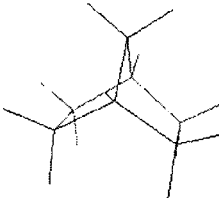
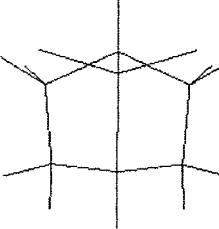
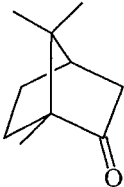
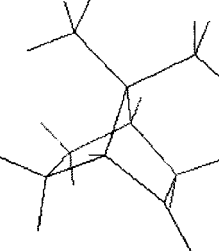
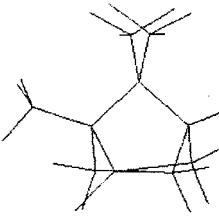


Figure 1. 2D representations of norbornane and camphor following adjustment of the weighting matrix.

Table 2. Results of a Straightforward NLM on Molecules of High Sphericity

Structural formula	Three dimensional representation	Two dimensional representation.	Sphericity (Ω)	Mean Error (\bar{A})
			0.82	0.42
			0.91	0.53

NLM takes between 50 and 200 cycles to reach the minimum error structure. As with all minimization algorithms we must consider the possibility that the system may get caught in a local, rather than global minimum. To test whether this was possible the algorithm was started from a number of randomly chosen positions for each structure illustrated. At all times the routine was found to converge to the same, low error, structure. On a PC486-33 this consumes approximately 2 min of CPU time for a 50 atom molecule. This compares unfavorably with the PCA analysis, which for all practical purposes is instantaneous. However as pointed out the NLM need not be started from random positions. Starting the NLM with the positions generated by PCA yields massively reduced iteration times, about 10–20 iterations being all that are required to reach the same lower error structure. Clearly this is considerably faster, making the NLM an insignificant consumer of CPU time. This illustrates an important point, the NLM should not to be considered to be competing with PCA, instead they should be utilized in a complementary fashion, PCA lending the NLM speed, the NLM lending PCA accuracy.

Finally, PCA is very much a “one shot” technique, which of course explains its speed. However this is not always desirable. Traditionally 3D to 2D mapping algorithms, like PCA have had problems with molecules which are highly spherical. These problems are 2-fold. Firstly an essentially spherical structure will have a high error when converted into its 2D representation. Whilst, as before, the NLM is able to improve on the PCA generated structure somewhat,

the error is still larger than that associated with flatter structures. This is clearly a mathematical fact of life, rather than an explicit flaw in either technique. If we denote the eigenvalues of the auto-covariance matrix used in the PCA analysis of the molecule as λ_1 , λ_2 , λ_3 , then the following coefficient Ω , which we call a sphericity coefficient, provides a numerical warning of potential problems:

$$\Omega = \frac{3\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3}$$

Ω varies from 0 for totally flat molecules, such as benzene where errors are small, to unity for totally spherical molecules such as methane, where distortions are liable to be greater.

Provided that we are willing to live with the greater error in the 2D representations, which is often not too serious, we run into the second traditional limitation of 3D to 2D mapping algorithms when dealing with nearly spherical molecules. This concerns the consistency of the positioning of the principal plane. Table 2 shows the results of the NLM when applied to norbornane and camphor with all elements of the weighting matrix set to unity. As can be seen the 2D representations generated lack the comparability which is required for a successful and meaningful similarity calculation. However, consider Figure 1 which shows the results of the NLM on the same structure after the weighting matrix has been altered to lend greater emphasis on the common structure between the molecules. As can be seen, the 2D representations are now consistent with each other and should

present no problems when considering evaluating their similarity. In this case the weighting matrix was set so that those distances which had a greater error in the original 2D representation were lent greater weight in the second mapping process, a technique that did not require having a consistent numbering system between the norbornane and camphor molecules and which appears reasonably general.

CONCLUSIONS

We have shown in this paper that the NLM algorithm which we introduced has great promise when applied to smaller organic molecules. We have demonstrated that we can augment the NLM's ability to form 2D representations of low error with the speed of execution of more traditional techniques such as PCA. Finally we have shown that the NLM can be naturally extended to encompass problems which present serious problems for the more traditional techniques.

It is our firm belief that these 2D representations will be invaluable in speeding up the evaluation of quantitative

similarity measures for large batches of molecules. The results of some of our investigations into this application will be published in the following paper.

ACKNOWLEDGMENT

Thomas Barlow would like to thank Balliol College, Oxford for a Janssen Junior Research Fellowship. In addition we would like to thank the anonymous reviewers whose constructive criticisms were invaluable in bringing focus to this work.

REFERENCES AND NOTES

- (1) Barlow, T. W.; Richards, W. G. A Novel Representation of Protein Structure, *J. Molecular Graphics* **1995**, *13*, 373–376.
- (2) Sammon, J. W. A Nonlinear Mapping for Data Structure Analysis. *IEEE Trans. Comput.* **1969**, Vol-C18, No. 5, 401–409.
- (3) Concepts and Applications of Molecular Similarity; Johnson, M. A., Maggiora, G. M., Eds; Wiley-Interscience: New York,
- (4) Havel, T. F.; Kuntz, I. D.; Crippen, G. M. The theory and practice of distance geometry. *Bull. Math. Biol.* **1983**, *45*, 665–720.

CI970424L