# Similarity Calculations Using Two-Dimensional Molecular Representations

Benjamin C. P. Allen, Guy H. Grant, and W. Graham Richards*

New Chemistry Laboratory, University of Oxford, South Parks Road, Oxford OX1 3QT

Received July 29, 2000

Molecular similarity calculations are important for rational drug design. Time constraints prevent these techniques being used on large data sets or on large molecules. By reducing the molecular representation to a two-dimensional form, the alignment of the molecules can be greatly speeded up. The accuracy of the resulting similarity values can be improved by using a neural network.

## 1. INTRODUCTION

Molecular similarity is a measure of the degree of overlap between a pair of molecules in some property space. It can be calculated for a wide range of molecular properties, including shape, electron density, electrostatic potential, lipophilicity, and refractivity.[1] A number of different functional forms have been used to quantify these measurements, of which the earliest is the Carbo index.[2] Molecular similarity is an important tool in the generation of quantitative structure−activity relationships (QSARs), which aim to link systematically the chemical or biological properties of molecules to their structures and are widely used in rational drug design.[3] With the advent of high throughput synthesis and combinatorial chemistry techniques, extremely large sets of potentially bioactive molecules can be created. With such large amounts of data to handle, the efficient storage of data and the rapid computation of molecular similarity values become crucial. The use of two-dimensional molecular representations can help with both these issues. Additionally there is considerable interest in molecular similarity calculations applied to proteins, which are currently restricted by time constraints.

Dimensionality reduction was first investigated as a method of analyzing higher dimensional data sets. By reducing multivariate data to two dimensions, it becomes possible to identify clustering and patterns by eye.[4] The most common method is principal component analysis (PCA), which extracts a set of orthogonal variables that capture a maximum amount of the variation in the original data set.[5] PCA is computationally highly efficient, using virtually no CPU time to reduce a system of 50 datapoints from three to two dimensions. However it is overly sensitive to outlying data points or badly distributed data and does not perform well on molecular structure data.[6] Additionally it is a one-shot technique, which means that when dealing with highly nonplanar molecules the results it produces are often heavily dependent on the postitioning of the principal plane, and it is thus hard to make comparisons between molecules. As a result, work on dimensionality reduction of molecular systems has concentrated on nonlinear methods, such as Kohonen networks[7] and Sammon mapping.[4]

Kohonen mapping can be used to produce topology preserving maps of molecular surfaces, and these have been used to compare the electrostatic potential surfaces of several classes of molecules.[8] It was shown that muscarinic and nicotinic agonists could be distinguished by their potential maps. More quantitative results have been obtained by calculating similarities between such maps:[9] a set of histamine H2 receptor agonists were ranked according to the similarity of the two-dimensional maps of the electrostatic potential surface, and it was shown that the ranking agreed with the measured biological activities of the molecules.

This work uses Sammon mapping, a nonlinear mapping algorithm, which can produce a dimensionally reduced structure with minimized distance matrix errors.[4] This algorithm has been used to produce two-dimensional representations of small molecules[6] and of protein alpha carbon positions.[10] There are many techniques for the processing and manipulation of two-dimensional forms, mostly developed in the field of digital image processing. Three such techniques have been used to prealign these representations to perform molecular similarity calculations.[11] The techniques tested were as follows: invarient moments, where the molecule is characterized by a set of seven translation and rotation invarient moments and radial integration or scanning, where the molecule is centered in the coordinate system and some property is either integrated or scanned along a radius and the rotationally invarient correlation integral calculated. These methods were tested on a set of steroids, and correlations with three-dimensional similarity indices of 0.62, 0.73, and 0.74, respectively, were obtained for the three approaches.

The method developed here is a three-step process for the calculation of molecular similarity indices. First the target molecules are reduced to two-dimensional forms. Next the optimum value of the similarity index is obtained for each pair of molecules by systematic rotation in relative configurational space. This is the most commonly used technique for three-dimensional similarity calculations. For highly similar sets of molecules this method is more time-consuming than prealignment, but for diverse sets it provides much better results. Finally the resulting matrix of two-dimensional similarity values is optimized using a neural net.

* Corresponding author phone: (0)1865 275908; fax: (0)1865 275905; e-mail: graham.richards@chem.ox.ac.uk.

10.1021/ci0003956 CCC: $20.00 
Published on Web 03/03/2001

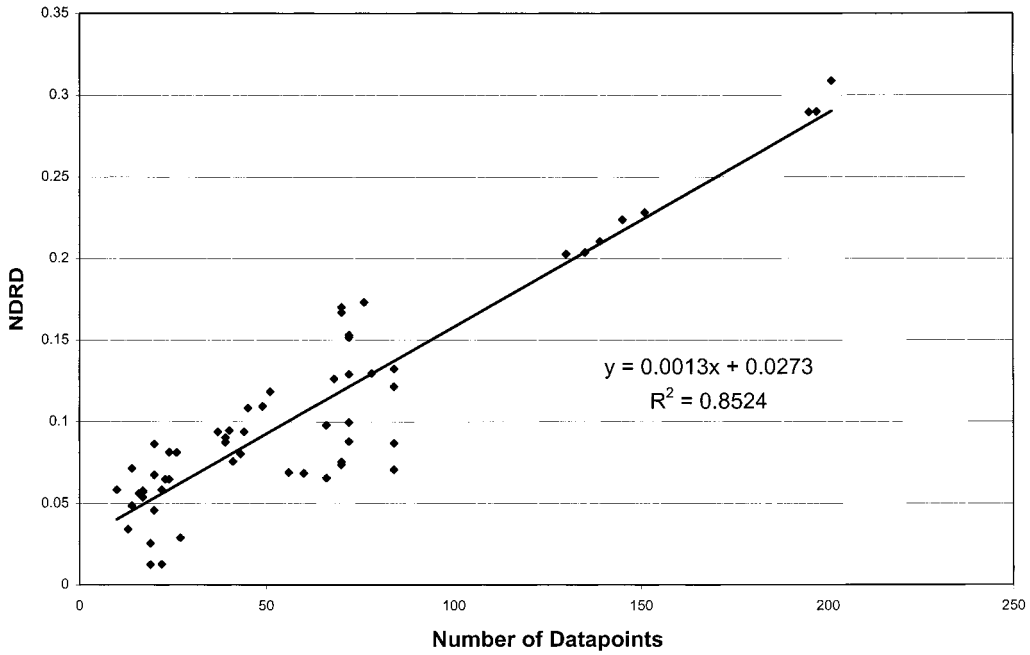



**Figure 1.** Accuracy of the Sammon mapping for systems of differing sizes at a runlength of 50 000 cycles.

## 2. METHODS

**2.1. Sammon Mapping.** Sammon mapping is a nonlinear algorithm that converts a *p*-dimensional set of data points into a corresponding *q*-dimensional set, where *p* is greater than *q*, while minimizing the differences between the distance matrices of the two sets.[4] In this work it is used to convert three-dimensional molecular structure coordinates into two-dimensional equivalents. To calculate these, consider a molecular structure consisting of *n* atomic positions in some three-dimensional reference frame:

$$\{\mathbf{A}_i\}_{i=1}^{n} \quad \text{where } \mathbf{A}_1 = [a_{1,1}, a_{1,2}, a_{1,3}] \tag{1}$$

An equivalent set of coordinates in a two-dimensional reference frame is constructed, by randomly selecting the initial positions

$$\{\mathbf{B}_i\}_{i=1}^{n} \quad \text{where } \mathbf{B}_1 = [b_{1,1}, b_{1,2}] \tag{2}$$

then the distance matrix is computed for each set

$$d_{ij}^{\mathbf{A}} = \left(\sum_{k=1}^{3}(a_{i,k} - a_{j,k})^2\right)^{1/2} \tag{3}$$

$$d_{ij}^{\mathbf{B}} = \left(\sum_{k=1}^{2}(b_{i,k} - b_{j,k})^2\right)^{1/2} \tag{4}$$

and the error matrix calculated

$$E_{ij} = \frac{1}{\sum_{i<j} d_{ij}^{\mathbf{A}}} \sum_{i<j} \frac{[d_{ij}^{\mathbf{A}} - d_{ij}^{\mathbf{B}}]^2}{d_{ij}^{\mathbf{A}}} \tag{5}$$

The application of Sammon mapping consists of minimizing the error matrix by adjusting the two-dimensional coordinates *b*, using the steepest descent algorithm.[4] The end set of two-dimensional coordinates will have a distance matrix that is maximally similar to that of the three-dimensional coordinates.

The accuracy and computational efficiency of the method were tested using five sets of different sized objects: amino acids (10−30 atoms), DHFR inhibitors (35−50 atoms), HIV protease inhibitors (50−70 atoms), HIV proteases (130−150 alpha carbons), and kinesins (190−200 alpha carbons), with the optimization parameter set to 50 000 cycles. The results are shown in Figures 1 and 2, with the errors calculated in terms of the normalized distance matrix root-mean-squared deviation (NDRD), given by

$$E_{nm} = \frac{1}{\sum_{i\neq j}^{N} d_{ij}} \left(\sum_{i\neq j}^{N} (d_{ij}^{n} - d_{ij}^{m})^2\right)^{1/2} \tag{6}$$

This form of error measurement was used because it produces a dimensionless error parameter allowing error comparisons between different sizes of molecules. A more standard error measure can be obtained by multiplying the NDRD by the average interatomic distance for the molecule. The errors increase linearly with the number of datapoints and remain quite small even for 200 datapoint systems. The time required to produce the two-dimensional representation is quadratic in the number of data points and remains reasonably small even for the largest systems tested. This indicates that the Sammon mapping is suitable for use in systems of up to 200 datapoints. Further speed increases can be obtained by using principal component analysis to produce a starting set of two-dimensional positions.[12]

The Sammon mapping was performed 10 times on the entire set of amino acids at each of five different optimization settings. The results, the average result at each setting and the average deviation from the mean of each set of results, are shown in Figure 3. They indicate that for a run length of 10 000 cycles the NDRD between the two- and three-

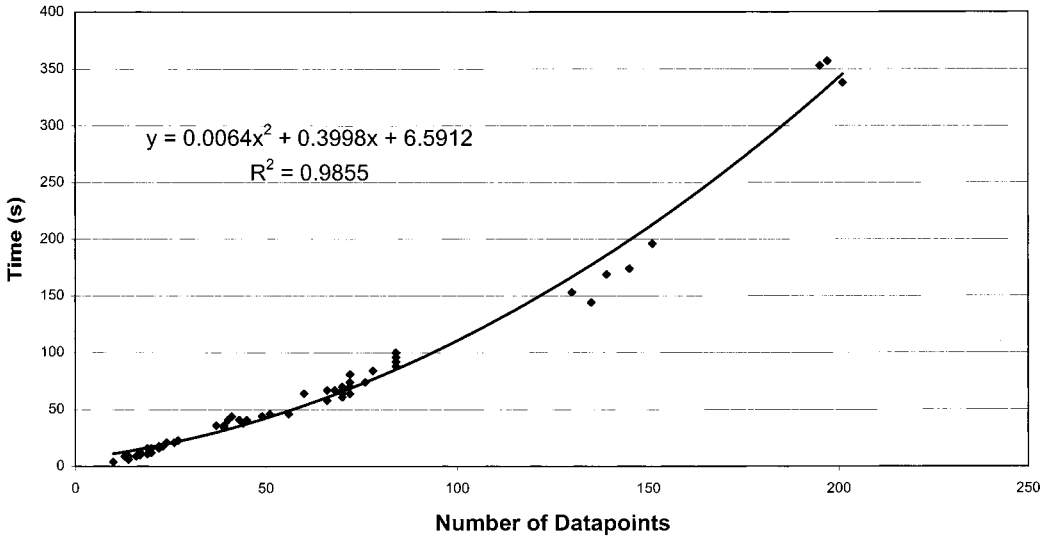



**Figure 2.** Computational efficiency of the Sammon mapping for systems of differing sizes at a runlength of 50 000 cycles.

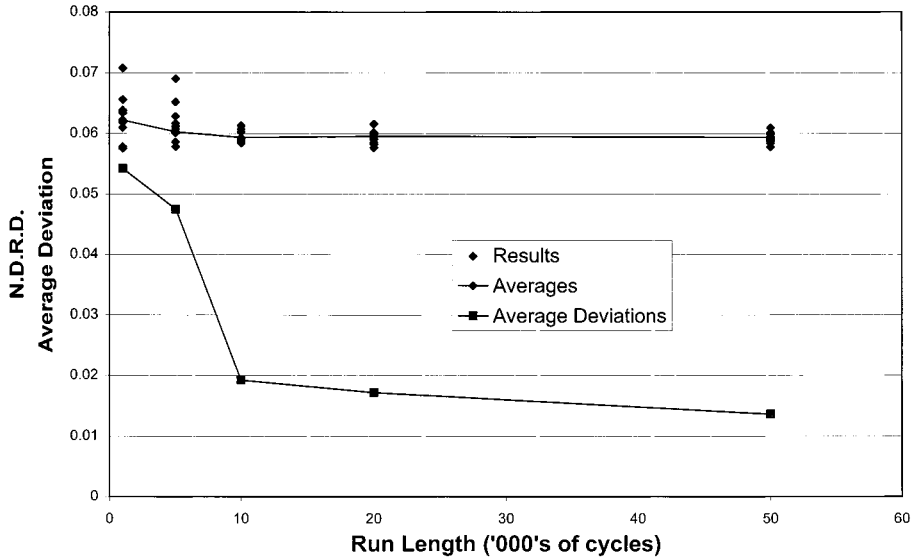



**Figure 3.** Accuracy and consistency of the results of the Sammon mapping for the amino acids at differing optimization times.

dimensional structures is 0.06 and that this value does not significantly improve for larger run lengths. The average deviation from the mean is 1.9%, while the value at 50 000 cycles is 1.5%, which indicates that all the runs are likely to have reached their final structures by 10 000 cycles. While it was not possible to prove formally that every system studied has a single minimum, the results at 50 000 cycles are so close to identical that it is clear that each structure is converging to the same minimum. For larger systems this was not the case, but it has been shown that for proteins the same structure is always found if the initial two-dimensional configuration has no crossings.[13] For more highly branched structures, the same general principle can be applied, and it was determined that, to ensure reproduceabilty, the two-dimensional structure should be initialized with as few bond crossovers as possible.

The Sammon code used was obtained from the SOM_PAK, a set of programs for the creation and testing of self-organizing maps.[14]

**2.2. Similarity Calculations.** Molecular similarity indices measure the degree of overlap of some spatially varying property between two superimposed molecules. They can be calculated using a wide range of molecular properties and by several techniques.[1] This work uses the Carbo index, which measures the degree of overlap of the electron density, $\rho$, given by

$$R_{AB} = \frac{\int \rho_A \rho_B d\upsilon}{(\int \rho_A^2 d\upsilon)^{1/2} (\int \rho_B^2 d\upsilon)^{1/2}} \quad (7)$$

The main term in the electron density function exhibits a $1/r$ dependence, and for computational efficiency this was approximated by a three term Gaussian function:[15]

$$\rho(r) \approx 0.3001e^{-0.0499r^2} + 0.9716e^{-0.5026r^2} + 0.1268e^{0.0026r^2} \quad (8)$$

Use of this approximation reduces the time taken to calculate similarities of presuperimposed pairs to the order of 1 s, even for large molecules.

Superposition of the two target molecules is the most time-consuming step in the calculations. The technique used involves fixing one molecule and moving the other in relative

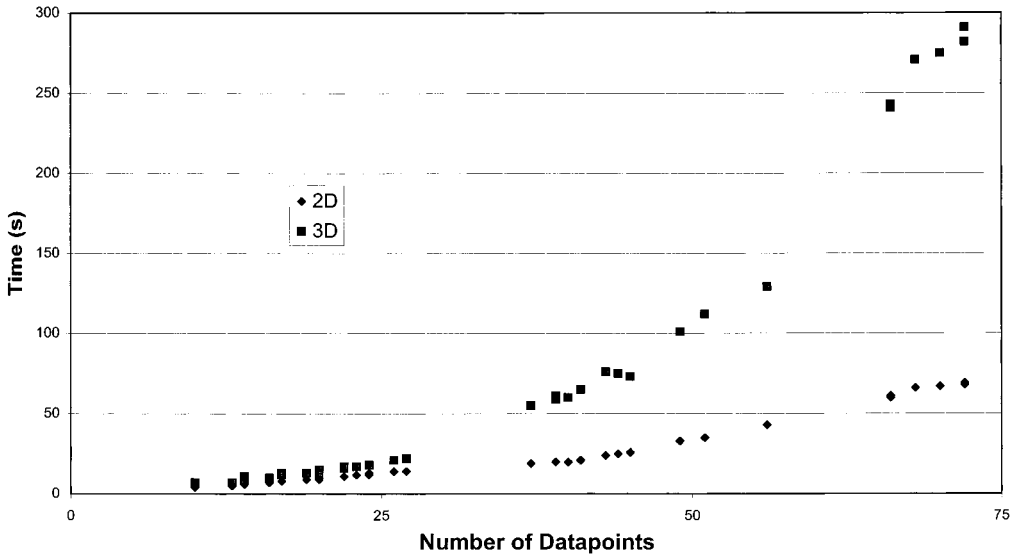



**Figure 4.** Superposition times for individual molecules, calculated by averaging the superposition times of the specified molecule with each of the other members of its set. The sets consist of the 20 amino acids (less than 30 datapoints), 10 DHFR inhibitors (35−60 datapoints), and 5 HIV protease inhibitors (over 65 data points).

configuration space to achieve maximum similarity. Ignoring conformational issues, for three-dimensional molecules the relative configurational space has five degrees of freedom, while for two-dimensional representations this reduces to three. The timings for this method were calculated for the 20 amino acids, 10 DHFR inhibitors, and 5 HIV protease inhibitors, using both the two- and three-dimensional representations of the molecules. The average time for comparing each molecule against all the others in its set is shown in Figure 4. It is clear that the use of two-dimensional representations significantly speeds up the superposition, especially for large molecules. In addition the use of two-dimensional representations allows techniques developed for pattern recognition to be employed, for even greater speed increases with highly similar data sets.[11]

The similarity calculations and superpositioning were carried out using the ASP module[16] within the program TSAR.[17]

**2.3. Neural Networks and Data Optimization.** Neural networks can be used to remove some fraction of the random error from a data set and show how closely the underlying data agree with some other data set.[18] This work used a set of properties obtainable from the two-dimensional representation to try and derive a modified similarity index as close as possible to the three-dimensional index. Several different properties were investigated, and a particular set of four was found to provide the most rapid and consistent results. A simpler neural network where no additional properties were used proved difficult to generalize. The network used was a standard back-propagation network with a sigmoidal transfer function, in which the network properties are determined by two parameters, learning rate, which controls how much the system changes per cycle, and the momentum term, which allows the system to escape from local minima.[18] The neural network was tested several times with different initialization settings, and it was shown that the network always converged to the same result for a given set of data, even with the momentum term set to zero. The optimal learning rate was determined to be a value of 0.35. A variety of network architectures were tested, and a simple structure with one layer of hidden units was found to be optimal. The number of hidden units was varied depending on the number of data points in the training set being used, so as to keep the critical ratio $\theta$ approximately equal to two. The critical ratio $\theta$ is given by

$$\theta = \frac{\text{no. of data points in the training set}}{\text{no. of variables in the network}} \qquad (9)$$

where the number of variables in the network is given by the total number of connections between units. It has been shown by many workers that a value of two provides the best predictive power.[19] The total run time for the network to generalize was found to be of the order of 20 s for 134 data points and to increase very slowly for larger systems.

The neural networks were created and tested using SNNS, the Stuttgart Neural Network System, a general purpose program for the creation and testing of a wide variety of network types.[19]

## 3. RESULTS

**3.1. Amino Acids.** This test set consisted of the 20 amino acids. The three-dimensional structures were constructed with ChemDraw and optimized using COSMIC, the optimization routine within TSAR.[17] They then underwent Sammon mapping to produce two-dimensional equivalent structures. The set of two-dimensional structures with the lowest NDRD provided the input for TSAR, and a complete set of 210 pairwise similarity values was produced using individual superposition prior to each calculation. This was compared with the corresponding set of three-dimensional similarity values, obtained using identical methods, giving a percentage correlation of 79%, calculated by Pearson's equation.[20] The similarity matrices for the two- and three-dimensional structures are shown in Figures 5 and 6, with the amino acids arrangement shown in Table 1. Some similarities between the two matrices can be observed, but there are numerous points of difference, principally due to single results that differ by substantial amounts.

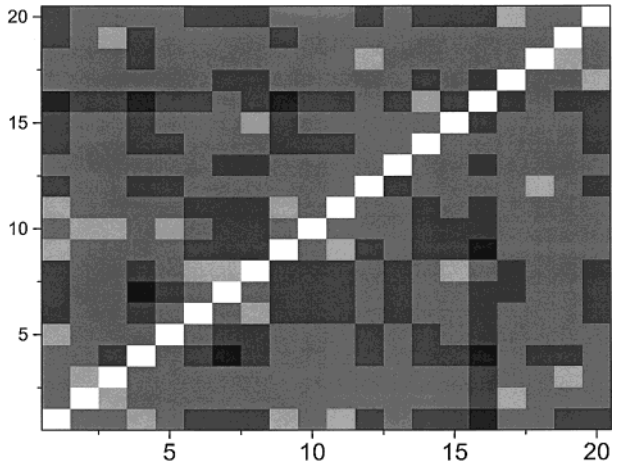



**Figure 5.** 3D amino acid similarity matrix.

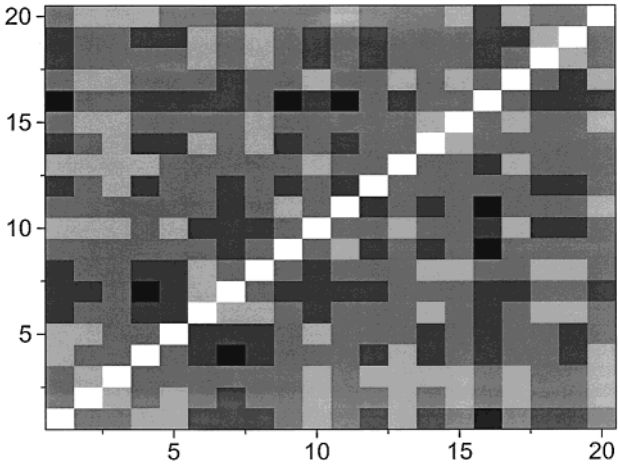



**Figure 6.** 2D amino acid similarity matrix.

**Table 1.** Amino Acid Arrangement

| | | | |
|---|---|---|---|
| 1 | alanine | 11 | cysteine |
| 2 | isoleucine | 12 | methionine |
| 3 | leucine | 13 | proline |
| 4 | glycine | 14 | lysine |
| 5 | valine | 15 | histidine |
| 6 | phenylalanine | 16 | arginine |
| 7 | tryptophan | 17 | aspartic acid |
| 8 | tyrosine | 18 | glutamic acid |
| 9 | serine | 19 | glutamine |
| 10 | threonine | 20 | asparagine |

The two-dimensional molecular structures were then used to calculate four additional properties for each molecule: the molecular mass, the total dipole moment, the Kappa 1 index,[21] and the Wiener topological index.[22] These four properties were selected because the values calculated from the two- and three-dimensional structures were similar or identical. The Kappa 1 index is given by

$$ {}^1\kappa = \frac{2\,{}^1P_{max}\,{}^1P_{min}}{({}^1P_i)} \tag{10} $$

where for any number of atoms $A$

$$ {}^1P_{max} = (A(A-1))/2 \tag{11} $$

$$ {}^1P_{min} = (A-1) \tag{12} $$

**Table 2.** Amino Acid Correlation Coefficients

| training | 1 | 2 | 3 |
|---|---|---|---|
| 2D | 0.797 | 0.798 | 0.801 |
| NN | 0.824 | 0.818 | 0.844 |
| test | 1 | 2 | 3 |
| 2D | 0.799 | 0.801 | 0.797 |
| NN | 0.863 | 0.859 | 0.895 |

and ${}^1P_i$ is the number of bonds in the structure, and it provides a measure of the complexity of the bonding pattern. The Wiener topological index is given by

$$ W = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} D_{ij} \tag{13} $$

with

$$ D_{ij} = \begin{cases} l_{ij} & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{14} $$

where $l_{ij}$ is the shortest distance between the atoms $i$ and $j$, and it measures the sum of terms in the distance matrix, providing a quantification of the degree of packing in the structure of the molecule. These four additional parameters provide a rough quantification of the shape of the molecule and indicate how well the Sammon mapping will perform.

Using these additional parameters a neural network was used to optimize the fit of the two- and three-dimensional similarity values. Three randomly selected nonoverlapping sets of 70 similarity values each were created and used to train neural nets, which were then applied to the other 140 values. The results are shown in Table 2. The three test and training sets all give fairly similar results, showing that the random selection procedure was sufficient. The overall correlation between the neural net optimized results and the three-dimensional values was 90%. Figure 7 shows the error profile of the two-dimensional similarity values with respect to the three-dimensional values for both the original and neural net optimized sets. The neural network reduces the mean error and also eliminates the outlying points of the error distribution, making the two-dimensional results a much more reliable approximation to the three-dimensional results. The amino acid set is sufficiently self-similar that the additional accuracy provided by the neural net is not particularly valuable. However the removal of points with very high error values makes the overall similarity between the data sets much more evident. Figure 8 shows the similarity matrix for the neural network adjusted two-dimensional similarity values, and the increased similarity between it and the 3D matrix is immediately evident, although a few minor features differ.

**3.2. HIV Protease Inhibitors.** This test set consisted of 119 cyclic urea derivatives, which have been shown to act as HIV protease inhibitors.[23] It differs from the amino acid set in having much larger structural differences between its members. The core structure and some of the possible substituents are shown in Figure 9. The set was created by attaching various groups to this core at the positions labeled $R_1$ and $R_2$, from a range of 90 substituents. The three-dimensional structures were obtained in prealigned form and

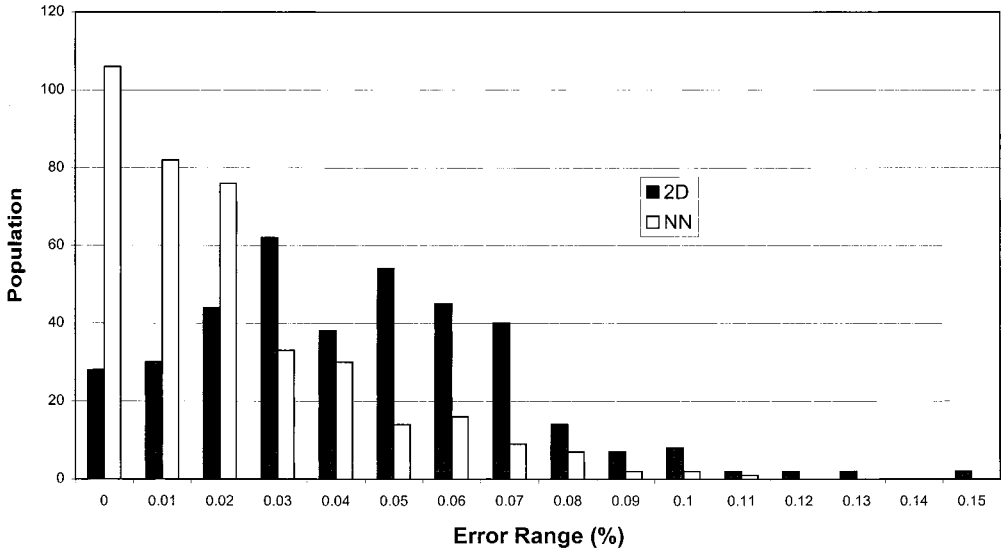



**Figure 7.** Error coefficients for the two-dimensional and neural net optimized two-dimensional amino acid similarity values as compared to the three-dimensional values.

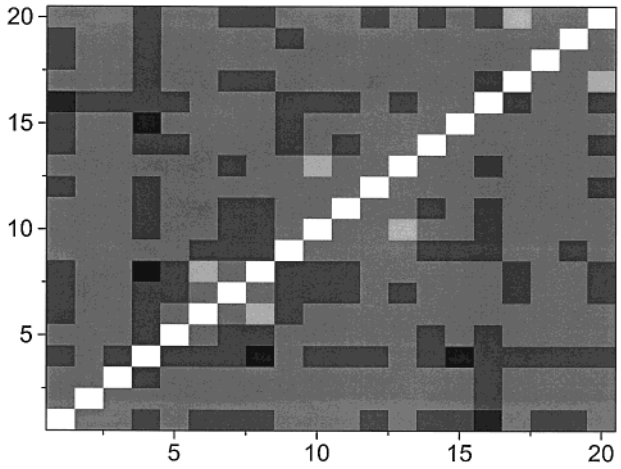
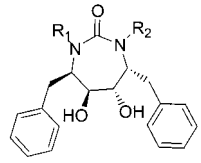



**Figure 8.** Neural net optimized amino acid similarity matrix.




Typical R groups include:

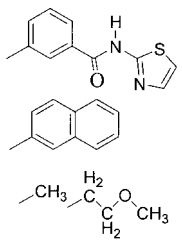



**Figure 9.** The HIV protease inhibitor core structure.

were then randomly rotated and realigned to provide timing data. The Sammon mapping was performed five times with an optimization parameter of 10 000 cycles, and the average deviation from the mean was calculated to be 2%. The set of two-dimensional structures with the lowest NDRD was used to produce complete sets of two-dimensional pairwise similarity values. These were compared with the three-dimensional values and a correlation of only 49% was obtained.

**Table 3.** HIV Protease Inhibitor Correlation Coefficients

| training | 1 | 2 | 3 |
|---|---|---|---|
| 2D | 0.467 | 0.454 | 0.516 |
| NN | 0.728 | 0.760 | 0.788 |
| test | 1 | 2 | 3 |
| 2D | 0.394 | 0.647 | 0.350 |
| NN | 0.796 | 0.854 | 0.755 |

As with the amino acid test set, four additional properties were calculated for each two-dimensional representation. The sets of similarity values for the molecules with the highest and lowest bioactivity (238 in total) were taken and randomly allocated to three nonoverlapping sets of 79, 79, and 80 values, respectively, which were used to train neural nets that were then applied to the corresponding test sets. The results for the three sets are shown in Table 3. It is clear that training set 2 provides a better representation of the whole set; however, the differences in results between the sets are acceptable. The overall correlation between the two- and three-dimensional results was calculated to be 86%. The error distribution is shown in Figure 10 and shows similar features to that of the amino acids. The neural net again does an excellent job of eliminating data points with very large error values.

## 4. DISCUSSION

The initial accuracy of the two-dimensional similarity values was shown to be very good for the set of 20 amino acids but quite poor for the set of 119 HIV protease inhibitors. The poor initial results for the inhibitor data set are due to the large size of the set and the wider range of structures, which gives rise to a much broader range of similarity values. This causes a larger average error, as the errors introduced by the dimensionality reduction process are much greater in dissimilar pairs of molecules. It has been shown that in both cases the accuracy of the two-dimensional results could be greatly increased by feeding the data into a

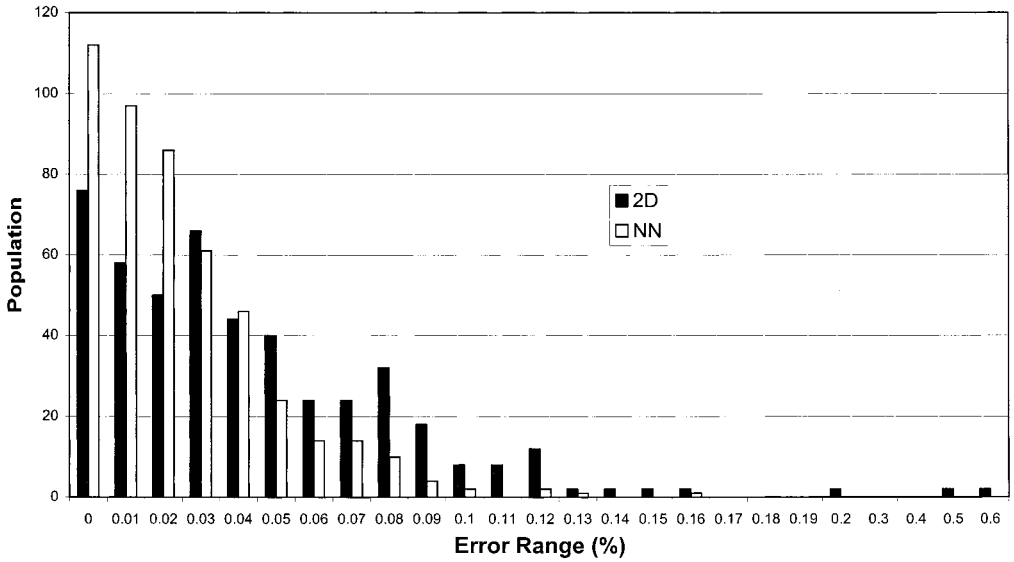



**Figure 10.** Error coefficients for the two-dimensional and neural net optimized two-dimensional HIV protease inhibitor similarity values as compared to the three-dimensional values.

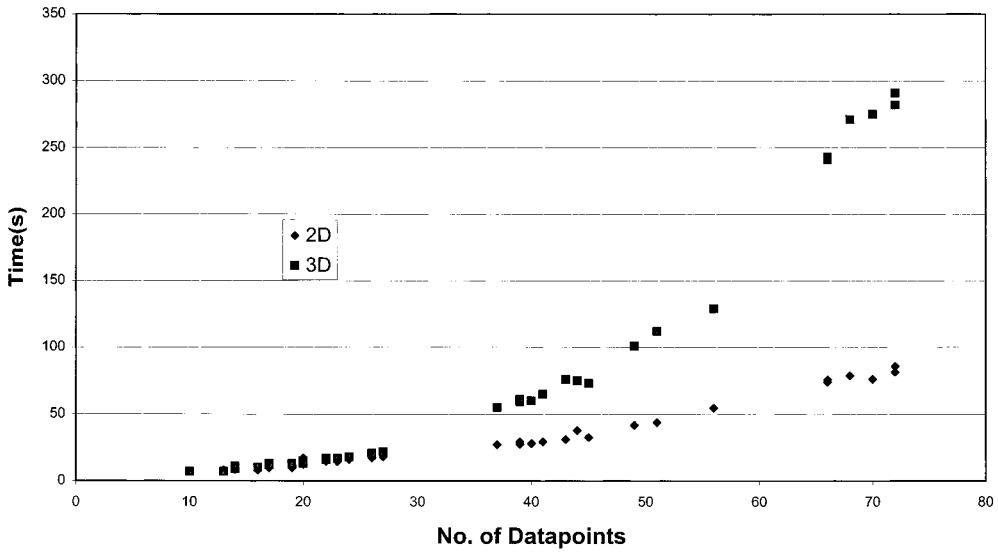



**Figure 11.** Comparison of total time requirements for two- and three-dimensional methods.

neural network. For the amino acid set, the improvement was statistically significant, but the similarity matrices indicate that even the unmodified two-dimensional values capture the main features of the similarity spectrum of the set. For the HIV protease inhibitor set, the neural net made a much greater contribution, improving the results from a very poor correlation of 49% to a significant correlation of 86%. It is worth noting that the correlation between different techniques for obtaining three-dimensional similarity values is generally of the order of 75%.

The results from this technique compare well with other methods based on dimensionality reduction. The techniques tested by Robinson et al. gave rise to similarity matrices with correlation with the three-dimensional results of less than 75%, for a set of 31 highly similar steroids.[11] Without neural net optimization, a correlation of 79% was obtained for the set of 20 amino acids, which is a slight improvement. For the much larger and more dissimilar inhibitor set, use of the neural net gave a correlation of 86%. The technique discussed here is capable of comparable results to other tested two-dimensional methods for small, similar data sets and can additionally provide very good results for complex data sets, which are much more susceptible to errors in the dimensionality reduction process. The overall timings for this method, the sum of the Sammon mapping time and the time taken to align the molecules, are compared with the corresponding three-dimensional timings in Figure 11. For molecules with over 20 atoms the use of the two-dimensional intermediates requires less computer time. Where a neural net is required, the total time for the method depends on the fraction of molecules used to create the neural net. For the inhibitor set, a value of 0.33 was used, and that gave a total timing less than half that required for the three-dimensional calculations. For larger sets the fraction can be reduced leading to significantly lower time requirements.

## 5. CONCLUSIONS

This work has shown that two-dimensional molecular representations can be produced using Sammon mapping, with distance matrix inaccuracies of the order of 10% of the average interatomic distance. Less than 10 s per molecule is required to produce these representations for systems with

fewer than 100 atoms. The time taken to superimpose these two-dimensional representations is significantly lower than that required for the three-dimensional forms. It is clear that for molecules with over 30 atoms the speed increase due to superimposing in two dimensions outweighs the time cost of Sammon mapping.

Overall, this technique appears to provide significant speed increases with minimal loss of accuracy for molecular similarity calculations on systems with more than 30 atoms.

## REFERENCES AND NOTES


(1) Good, A. C.; Richards, W. G. Explicit Calculation of 3D molecular similarity. *Perspectives Drug Discovery Design* **1998**, *9*−*11*, 321−338.

(2) Leyda, L.; Carbo, R.; Arnau, M. An electron density measure of the similarity between two compounds. *Intl. J. Quantum Chem.* **1980**, *17*, 1185−1189.

(3) Good, A. C.; So, S.; Richards, W. G. Structure−activity relationships from molecular similarity matrices. *J. Med. Chem.* **1993**, *36*, 433−438.

(4) Sammon, J. S. A nonlinear mapping for data stucture analysis. *IEEE Trans. Computers* **1969**, *C-18(5)*, 401−409.

(5) Cooley, W.; Lohnes, P. *Multivariate Data Analysis*; Wiley: New York, 1971.

(6) Barlow, T. W.; Robinson, D. D.; Richards, W. G. Reduced dimensional representations of molecular structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37(5)*, 939−942.

(7) Kohonen, T. Self-organised formation of topologically correct feature maps. *Biological Cybernetics* **1982**, *43*, 59−62.

(8) Rudolph, C.; Sadowski, J.; Gasteiger J.; Li, X.; Zupan, J. Representation of molecular electrostatic potentials by topological feature maps. *J. Am. Chem. Soc.* **1994**, *116*, 4608−4620.

(9) Barlow, T. W. Self-organising maps and molecular similarity. *J. Mol. Graph.* **1995**, *13(1)*, 24−27.

(10) Barlow, T. W.; Richards, W. G. A novel representation of protein structure. *J. Mol. Graph.* **1995**, *13(6)*, 373.

(11) Barlow, T. W.; Robinson, D. D.; Richards, W. G. The utilization of reduced dimensional representations of molecular structure for rapid molecular similarity calculations. *J. Chem. Inf. Comput. Sci.* **1997**, *37(5)*, 943−950.

(12) Lyne, P. D.; Robinson, D. D.; Richards, W. G. Alignment of 3d-structures by the method of 2d-projections. *J. Chem. Inf. Comput. Sci.* **1999**, *39(3)*, 594−600.

(13) Barlow, T. W. Reduced Dimensionality in Molecular Representation. Ph.D. Thesis, Oxford, 1996.

(14) Kangas, J.; Kohonen, T.; Hynninen, J.; Laaksonen, J. *SOM_PAK: The Self-Organising Map Program Package*; Technical Report A30; Helsinki University of Technology: 1996.

(15) Good, A. C.; Richards, W. G. Rapid evaluation of shape similarity using Gaussian functions. *J. Chem. Inf. Comput. Sci.* **1993**, *33(1)*, 112−116.

(16) A. S. P. *Automated Similarity Package. V3.11 Release Notes*; Oxford Molecular Limited: The Magdalen Centre, Oxford Science Park, Sandford on Thames, Oxford OX4 4GA, United Kingdom.

(17) T. S. A. R. *Tools for Structure Activity Relationships. Reference Guide 3.2;* Oxford Molecular Limited: The Magdalen Centre, Oxford Science Park, Sandford on Thames, Oxford OX4 4GA, United Kingdom.

(18) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists*; VCH Publishers: 1993.

(19) Zell, A. et al. *Stuttgart Neural Network Simulator. Users Manual Version 4.2*; University of Stuttgart, Institute for Parallel and Distributed High Performance Systems, and Unversity of Tubingen, Wihelm-Schickard-Institute for Computer Science.

(20) Harper, W. M. *Statistics*; Pitman Publishing: 1991; Chapter 14, pp 172−173.

(21) Hall, L. H.; Kier, L. B. The molecular connectivity chi indexes and kappa shape indexes in structure−property modeling. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; 1992; Chapter 9, p 367

(22) Wiener, H. Correlation of heats of isomerisation, and differences in heats of vapourisation of isomers, among the paraffin hydrocarbons. *J. Am. Chem. Soc.* **1947**, *69*, 2636.

(23) Debnath, A. K. Three-dimensional quantitative structure analysis relationship study on cyclic urea derivatives as hiv-1 protease inhibitors: Applications of comparative molecular field analysis. *J. Med. Chem.* **1999**, *42*, 249−259.


CI0003956