Computer Recognition of Activity Class from Molecular Transforms

Sir:

We wish to report the first successful application of computer pattern recognition to molecular transforms derived from x-ray diffraction data. Hitherto, pattern recognition studies of compound activity (as opposed to compound identity) have been based on molecular fragment coding as the source of structure description.[1,2] Although successful activity recognition from multidimensional structure descriptions suggests the possibility of extracting subtle structure–activity relationships, work based on the use of fragment coding as the source of structural descriptors has drawn criticism.[3] The difficulty is that, by describing a molecule in terms of a predetermined list of molecular attributes, such as molecular weight, ring size, presence of carbonyl, and so forth, one may prejudice the possible results and thus draw misleading conclusions regarding which structural moieties are significant in producing activity.

Molecular transforms[4] are computed directly from three-dimensional atomic coordinates and thus partake of none of the arbitrariness of fragment coding. For class recognition where molecular shape is important, molecular transforms provide a direct source of data embodying that shape. We have used a Wierl-type expression to calculate molecular transforms:

$$I(s) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} Z_i Z_j \frac{\sin (sr_{ij})}{sr_{ij}}$$

Here, Z_i is the atomic number of the ith atom in an N-atom molecule, and r_{ij} is the distance between atom i and atom j. The transforms were computed at 100 equi-spaced points in the range 1 $Å^{-1} \leq s \leq 31$ $Å^{-1}$ and reduced to 100-dimensional binary patterns by tabulating "1" 's in intervals where a zero-crossing (i.e., $I(s) = 0$) occurred and "0" 's elsewhere. A typical pattern contained about 20 zero-crossings in the s-range studied. The use of zero-crossings to characterize the molecular transforms in binary form is analogous to peak/no-peak coding of NMR or mass spectra.[5] For the relatively broad peaks encountered in a molecular transform, determination of the peak positions is more difficult than for NMR or mass spectral peaks, whereas the zero-crossings of the molecular transform are easily identified.

Our preliminary results are reported for a set of 89 compounds, of which 42 are known to be sedatives and 47 are known to be tranquillizers.[6] These are a subset of the 219 compounds used by Stuper and Jurs[1] and include phenothiazines, benzodiazepines, barbiturates, carbamates, carbinols, and aliphatic amides. The molecular transforms for 13 of these compounds were computed from three-dimensional coordinates obtained from the Cambridge Crystallographic Data Centre data file[7] (Summer 1975 update) via the National Institutes of Health (Division of Computer Research and Technology) Chemical Information System. For those compounds not present in the file, coordinates were generated from those of similar molecules with the use of a modification of the program ATCOOR[8] and standard bond lengths, angles, and torsion angles. The molecular transforms were reduced to binary patterns, as described above, and these patterns were used to train binary pattern classifiers by the error-correction-feedback method which has been described elsewhere.[9,10] Although a data base of 89 patterns does not justify the use of 100 features, feature elimination by the weight-sign method[9,10] reduced the dimensionality to 25, which is an acceptable value for this size data base.[11]

Using molecular transforms in this manner, we have produced: (1) a 25-component weight vector, trained on all 89 compounds, capable of correctly classifying all 42 sedatives and all 46 tranquillizers; (2) a 21-component weight vector, trained on 79 compounds, capable of correctly classifying 9 out of 10 tranquillizers not present in the training set; (3) a 25-component weight vector, trained on 79 compounds, capable of correctly classifying 10 out of 10 sedatives not present in the training set; (4) a 21-component weight vector, trained on 79 compounds, capable of correctly classifying 5 tranquillizers and 4 sedatives out of 5 tranquillizers and 5 sedatives not present in the training set. These findings demonstrate that molecular transforms, prepared and sampled in the manner we have described, contain sufficient information to permit machine recognition of these two activity classes.

It should be pointed out that, since the molecular transform as we have computed it depends only on interatomic distances and not on absolute atomic positions, it may be possible to use x-ray structure-factor data directly to compute the molecular transform. The F^2_{obsd} would be used to compute a Patterson function,[12] and the height and position of the Patterson peaks would supply the Z_iZ_j and r_{ij} information needed for the molecular transform computation. The potential difficulty would be the possible inclusion of some intermolecular distances, but the prospect of achieving activity classification of a substance directly from diffraction data prior to structure solution seems worth pursuing.

Although "cranking out structures" is generally frowned upon by crystallographers, our findings argue for the collection of extensive files of three-dimensional atomic coordinates for major compound classes to be used in computer-aided investigations of structure–activity correlations. The current state of the art in x-ray structure determination and the existence of the Cambridge Crystallographic Data Centre project make this a realistic goal.

Acknowledgment. This research was supported by the National Science Foundation under Grant MPS-74-01249.

References and Notes

(1) A. J. Stuper and P. C. Jurs, *J. Am. Chem. Soc.*, **97**, 182 (1975).
(2) B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, **96**, 916 (1974).
(3) R. J. Mathews, *J. Am. Chem. Soc.*, **97**, 935 (1975).
(4) H. Lipson and C. A. Taylor, "Fourier Transforms and X-Ray Diffraction", G. Bell and Sons, Ltd., London, 1958, p 9.
(5) H. B. Woodruff, G. L. Ritter, S. R. Lowry, and T. L. Isenhour, *Appl. Spectrosc.*, **30**, 213 (1976).
(6) E. Usdin and D. H. Efron, "Psychotropic Drugs and Related Compounds", 2d ed, DHEW Publication No. (HSM)72-9074, 1972.
(7) O. Kennard, D. Watson, F. Allen, W. Motherwell, W. Town, and J. Rodgers, *Chem. Brit.*, **11**, 213 (1975).
(8) J. E. Nordlander, Quantum Chemistry Program Exchange, Program No. 250.
(9) T. L. Isenhour and P. C. Jurs, *Anal. Chem.*, **43**, 20A (1971).
(10) C. L. Wilkins, R. C. Williams, T. R. Brunner, and P. J. McCombie, *J. Am. Chem. Soc.*, **96**, 4182 (1974).
(11) R. O. Duda and P. E. Hart, "Pattern Classification and Scene Analysis", Wiley, New York, N.Y., 1973, pp 69–70.
(12) See, for example, G. H. Stout and L. H. Jensen, "X-Ray Structure Determination", Macmillan, New York, N.Y., 1968, pp 260, 270. The Patterson function is a density function $\rho(xyz)$ which has maxima at points corresponding to interatomic vectors r_{ij}; the amplitudes of the Patterson maxima are proportional to the product of the atomic numbers Z_iZ_j of the corresponding atoms.
(13) On sabbatical leave from Simmons College, Boston, Mass.

Leonard J. Soltzberg,[13] Charles L. Wilkins*
Department of Chemistry, University of Nebraska-Lincoln
Lincoln, Nebraska 68588
Received January 26, 1976

Electron Spin Resonance Spectrum of the Perfluorocyclobutane Radical Anion[1]

Sir:

Perfluorocycloalkanes are known to form negative ions by electron attachment in the gas phase.[2] Also, chemical studies have shown that these molecules compete very effectively against nitrous oxide for the electrons released in the radiolysis