Chemosphere, Vol.17, No.8, pp 1617-1630, 1988
Printed in Great Britain

0045-6535/88 $3.00 + .00
Pergamon Press plc

On the Intrinsic Dimensionality of Chemical Structure Space

by

G.D. Veith, B. Greenwood[1], R.S. Hunter[2], G.J. Niemi, and R.R. Regal[3]

Environmental Research Laboratory-Duluth
6201 Congdon Boulevard
Duluth, MN 55804

[1]Computer Sciences Corporation
Falls Church, VA

[2] Center for Data Systems and Analysis
Montana State University
Bozeman, MT 59717

[3] Department of Mathematical Sciences
University of Minnesota-Duluth
Duluth, MN 55812

INTRODUCTION

An important expectation in chemistry and pharmacology is that similar chemical structures have similar properties and behavior. New industrial chemicals, pesticides, and therapeutics are often subtle modifications of "lead" structures with known chemical behavior. Chemical properties and reaction rates can be estimated from suitable homologs[1,2]. Moreover, the safety of untested chemicals is often evaluated by comparing the chemical to analogous structures for which toxicological data are available. Despite the widespread use of terms such as "homolog" and "analogs" in research, chemical similarity has evaded quantitative interpretation from a perspective where all chemicals are considered simultaneously. One reason is that chemical similarity is inherently a multivariate problem or, in other words, chemicals are simultaneously similar and different from many perspectives. We have approached chemical similarity by attempting to define a structure space in which all chemicals can be identified. Because there are so many potentially important variables, multivariate tools are necessary to reduce the dimensionality of this problem. When this is accomplished, we need to comprehend what this space means and what can be predicted from it. This paper is one of the first attempts to define chemical structure space for a large universe of chemicals.

The dimensionality and scaling of chemical structure space has been sought using thermodynamic properties and the biological activity of molecules[3,4,5,6]. In these approaches, data are systematically compiled for many chemicals, and information is transformed to a smaller set of variables, or principal components which account for most of the variation in the set. The principal components are used as orthogonal coordinates in a structure space and distances between chemicals in this space are used as a measure of similarity for cluster analysis of other multivariate techniques. While this approach may be an improvement to intuitive selection of analogs, it is limited by the practical problem that systematic data sets of essential chemical properties are available for a relatively small number of chemicals. A small number of points in a multivariate space precludes exploring modeling and validating complex phenomena. Multivariate analyses

can be accurate only if chemicals representing all of the diversity likely to be encountered are included in the data set from which the structure space is derived. One of the best ways to produce a stable space is to compile a large, representative chemical data set. Otherwise, the stability of the system may change if new kinds of structures are added.

To explore the intrinsic dimensionality of chemical structure space, we have selected a set of 19,972 chemical structures from registries of chemical production. Because data on chemical properties are available for less than one percent of the chemicals, we have turned to methods of quantitating structural variations derived from graph theory. A set of more than 90 graph-theoretic indices have been systematically computed for each of the 19,972 chemicals, and the dimensionality of the data set has been reduced to a set of eight principal components. Computer programs were developed for minicomputers to graphically display the "universe" of chemical structures through user-selected windows.

METHODS

Molecular topology treats a chemical structure as a group of vertices (atoms) connected by edges (bonds)[7]. Methods for computing molecular connectivity indices from chemical sub-graphs have been discussed at length[8,9,10] and will not be discussed in detail here. The indices are classified as framework, bond, and valence indices. Framework indices are derived from structures reduced to only carbon atoms and single bonds. The bond indices provides a mechanism to look a step beyond framework indices in that all the vertices are assumed to be carbon and the vertex corrections differentiate the local bonding of each vertex. The correction factor for valence is the number of non-hydrogen bonds at the vertices. The valence indices use vertex values which are adjusted for both bonding and heteroatom electronegativity[9].

A graph is a finite set of vertices and a finite set of edges in which edges connect two of the vertices. A connected subgraph of a graph is a subgraph that has all the vertices connected by some combinations of the edges. Subgraphs are classified into paths (-C-C-), clusters (-C-C(-C)-C-), path-clusters (C-C-C(-C)-C) and cycles[9] (-C-C(-C)-C). A path is a non-cyclic subgraph that has only one or two edges to each vertex. A cluster is a non-cyclic subgraph that has only three or four edges to each vertex. A path/cluster is a non-cyclic subgraph that is composed of both a path and a cluster. A subgraph that contains at least one cyclic subgraph is defined as a chain. The order of a subgraph is the number of edges in the subgraph.

Indices of low order can be generated by hand calculation. However, systematic calculation of higher order indices for multicyclic molecules has not been reported due to the difficulties of accurate subgraph enumeration. We developed an efficient algorithm to compute the first 10 orders of indices using computer data structures of only connected subgraphs. The program efficiently uses identified subgraphs to generate new subgraphs which includes adjacent vertices and the indices are computed by simple bookkeeping of vertex type and number of edges at each vertex.

The graph enumeration program was developed on a VAX-11/780 computer at Montana State University. In an effort to gain more insight into the nature of molecular connectivities indices, particularly as a tool to determine structural and chemical similarity in molecules, we generated the connectivity indices for 19,972 chemicals selected from the U.S. EPA Toxic Substances Initial Inventory. The selected data base includes only discrete organic molecules with less than 60 non-hydrogen atoms and at least one carbon atom. Generating all indices for these chemicals took approximately 20 hours of CPU computer time on the VAX-11/780.

The 0th to 9th order terms for paths, the 3rd to 9th order terms for clusters, the 4th to 9th order terms for path/clusters, and the 3rd to 9th order terms for cycles for the framework, bond, and valence indices comprise 90 structural variables. Principal component analysis (PCA) was used to

explore the covariance structure of these variables and to reduce them to a set of orthogonal variables (principal components) that still retained a large part of the variation in the original connectivity indices. We calculated the principal components from the correlation matrix derived from the 90 variables calculated for 19,972 chemicals using the Statistical Package for the Social Sciences[11]. Because all the variables were skewed due to the presence of some large molecules relative to the majority of the chemicals in the data set, the variables were log-transformed to stabilize the variation and reduce the influence of these large molecules in the principal component space.

Designing a computer generated display of a chemical structure-space defined by 19,972 data points in many dimensions is a challenge. We wanted to display as much information in as many dimensions as possible. We began exploring the data set (hereafter termed the universe) through an expanded window of a three-dimensional spatial representation of the data and scaling a fourth dimension over a highly resolved color lookup table. To distinguish patterns further, rotation and magnification capabilities were also developed so we were able to "zoom in" for a closer examination at specific segments of the universe from different angles and dimensions.

Five years ago the cost of such a computer system would have been in excess of $100,000. Today, several sophisticated medium-to-high resolution color graphics devices are priced below $5000. The images presented herein were created on a Vectrix VX384 graphics device with a IBM PC AT host. The system uses an Intel 8088 microprocessor in conjunction with a NEC 7220 graphics device controller chip driving nine bit-planes of display memory, and is capable of displaying 512 colors. Each bit-plane is a cartesian coordinate grid consisting of 672 x 480 bits, each bit representing the spatial position of a dot on the screen. By scanning all nine planes at a particular coordinate, a nine-bit address is obtained in the color lookup table where a numerical setting for each color gun is stored. These settings determine what color is displayed at that coordinate on the grid. Each color gun on the monitor is resolved to eight bits or 256 parts. This provides a palette of 16.8 million choices for the 512 addresses of the color lookup

table. The three-dimensional transformations, rotations, and magnifications are all firm-ware implementations which greatly reduce the computational complexity of the program running on the host computer. This means sophisticated graphics programs that once could be run only on large mainframes are now within the capabilities of smaller mini- and microcomputer systems.

Three-dimensional windowing was implemented in the graphics primitives of the binary driver on the host IBM PC AT. A program called PROPCA and its associated routines were written in FORTRAN 77 to provide the benefits of a broad base of software compatibility and reasonably fast execution times. PROPCA allows the user to assign any three variables to the X, Y, and Z axes. A fourth variable is then selected to be mapped over the 512 colors of a linear spectral lookup table. A three-dimensional virtual window can be defined in terms of a minimum and maximum for the X, Y, and Z axes. This feature allows close examination of small sections of the image that are not available through the scaling options of the program. For example, one could select different ranges and endpoints for each axis. Also plot time is substantially reduced by selectively viewing only the areas of interest through the windowing option.

RESULTS

Principal components from the 19,972 x 90 data matrix were used as variables in this study and were retrieved from rapid traversals of an inverted file in the host computer. The principal components were plotted as single pixels as they were read. Figure 1 presents a first glimpse of the chemical universe and structure-space. A reference wire-frame cube describes the de fined virtual-window, providing orientation when rotating or scaling. A counter is also provided to inform the user of how far the plotting operation has progressed in terms of the data set traversal. After the traversal is complete, the counter is updated to display the number of points within the window limits. A color legend is displayed at the far right of the screen which allows subtle hues to be correlated with component values. Because the use of single pixels rather than filled polygons lessens the

capture of depth along the Z axis, PROPCA provides for projections onto the sides of the user-defined virtual window and concurrent viewing of the spatial and color information in the X-Y, X-Z and Y-Z planes.

The PCA resulted in eight principal components with eigenvalues > 1 and they explained 93.5% of the variation in the original data (Table 1). PC 1 was positively correlated with all variables except for the cyclic variables. PC 2 was positively correlated with all cluster variables that indicate the degree a molecule is branched, but negatively correlated with all path and cyclic variables. In contrast, PC 3 was positively correlated with all cyclic variables and negatively correlated with all other variables except the valence-corrected cluster and path/cluster variables. The first three principal components all convey generalized information on chemical structure: size (PC 1), degree of branchness (PC 2), and number of cycles (PC 3). The remaining five principal components identified more specific differences between chemicals. For example, PC 4 had positive correlations (r > .58) with the 3rd and 4th order cyclic variables, but negative correlations (r < .18) with the 7th and 9th order cyclic variables. Similarly PC 5 to PC 8 convey additional differences in branching, bonding, cyclicness, valency (presence of heteroatoms such as halogens and oxygen), and combinations of these structural attributes.

Figure 1 presents the chemical structure-space for the first three principal components on the X, Y, and Z axes, respectively. Color scales the fourth principal component with red to designate small values and blue to designate large values. The principal components are axes that represent gradients of differences between chemicals. For example, on the extreme left in Figure 1 is carbon monoxide, the smallest molecule in the data base, while the largest molecule in the data base (CAS # 1356089) has the largest value for PC 1. The "string" of structures or linear cluster in the lower left corner of Figure 1 is a group of nearly 1200 unbranched, non-cyclic structures which are separated from the universe of branched structures. Figure 2 is a different view of these same sets of principal components. Both views illustrate that structures which are close (similar) to each other in three dimensions may actually be far apart (dissimilar) in a fourth

dimension. Figure 3 presents another view of the universe from the perspective where the 4th, 5th, and 6th principal components present the X, Y, and Z axes respectively and the 7th principal component are scaled in color. This view contains 19,584 structures and many homologous series of chemicals are apparent in these dimensions.

DISCUSSION

This approach to developing a stable, multivariate definition of chemical similarity is being used for two purposes. The first is to identify suitable analogs defined by nearness in eight dimensions. It is beyond the scope of this paper to present possible algorithms for measuring distances. Nonetheless, using a Euclidean distance measure and eight principal components, we can report that if a molecule such as diphenylamine is inserted in the universe, its nearest neighbors include many of the 4,4'-hydroxyamino, hydroxyl, and methylamino derivatives of biphenyl. If phenoxymethyl oxirane (phenyl glycidal ether) is inserted, the nearest neighbors include 2-methyl, 2-ethyl, and 4-nitro phenoxymethyl oxirane.

The second use is to attempt to identify potentially harmful chemicals by association with chemicals of known chemical or biological behavior. For example, Veith et al.[12] demonstrated that the highly bioaccumulative chemicals in food chains have a large n-octanol/water partition coefficient (Log P). Chemicals with Log P values greater than 4.0 are considered to have substantial bioaccumulation potential. Figure 4 presents a close-up of Figure 1 in which the color axis has been scaled to Log P instead of to PC 4. The figure illustrates that for many chemical classes the Log P value increases with molecular weight and/or molecular volume (X axis). Therefore, the red regions of the universe contains chemicals with large Log P[13] and new structures which fall in these areas could reasonably be presumed to have substantial bioaccumulation potential even if the Log P value is unknown. Figure 4 also shows that there are large molecules with low Log P values (blue data at right-center). Even though these structures appear immersed in other structures, they are widely separate in other dimensions and constitute large but non-accumulative chemicals such as sulfonic acid and azo dyes.

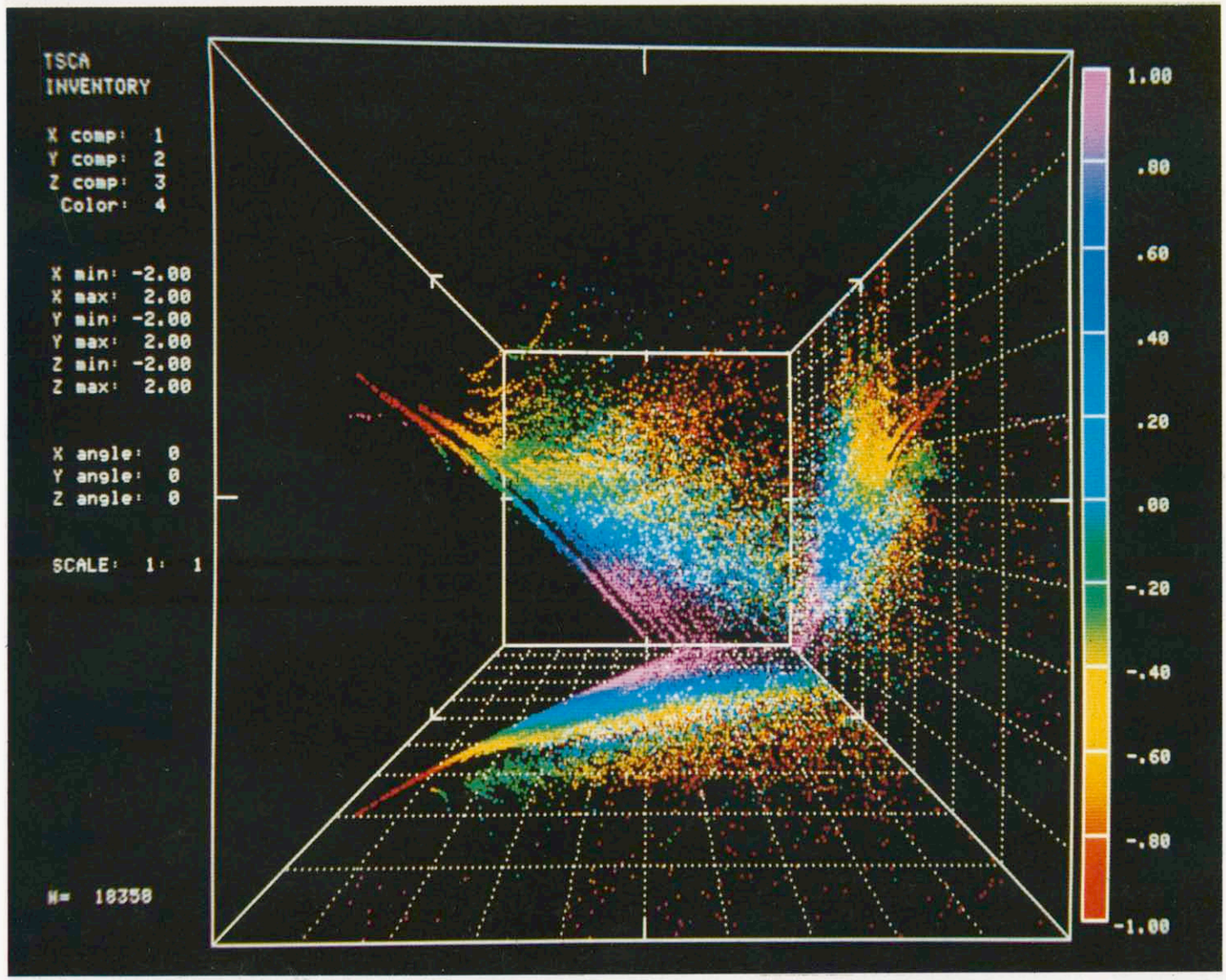

Figure 1. Graphical representation of the industrial chemical universe in structure space. The X, Y, Z and color axes are the first, second, third and fourth dimensions (principal components), respectively.

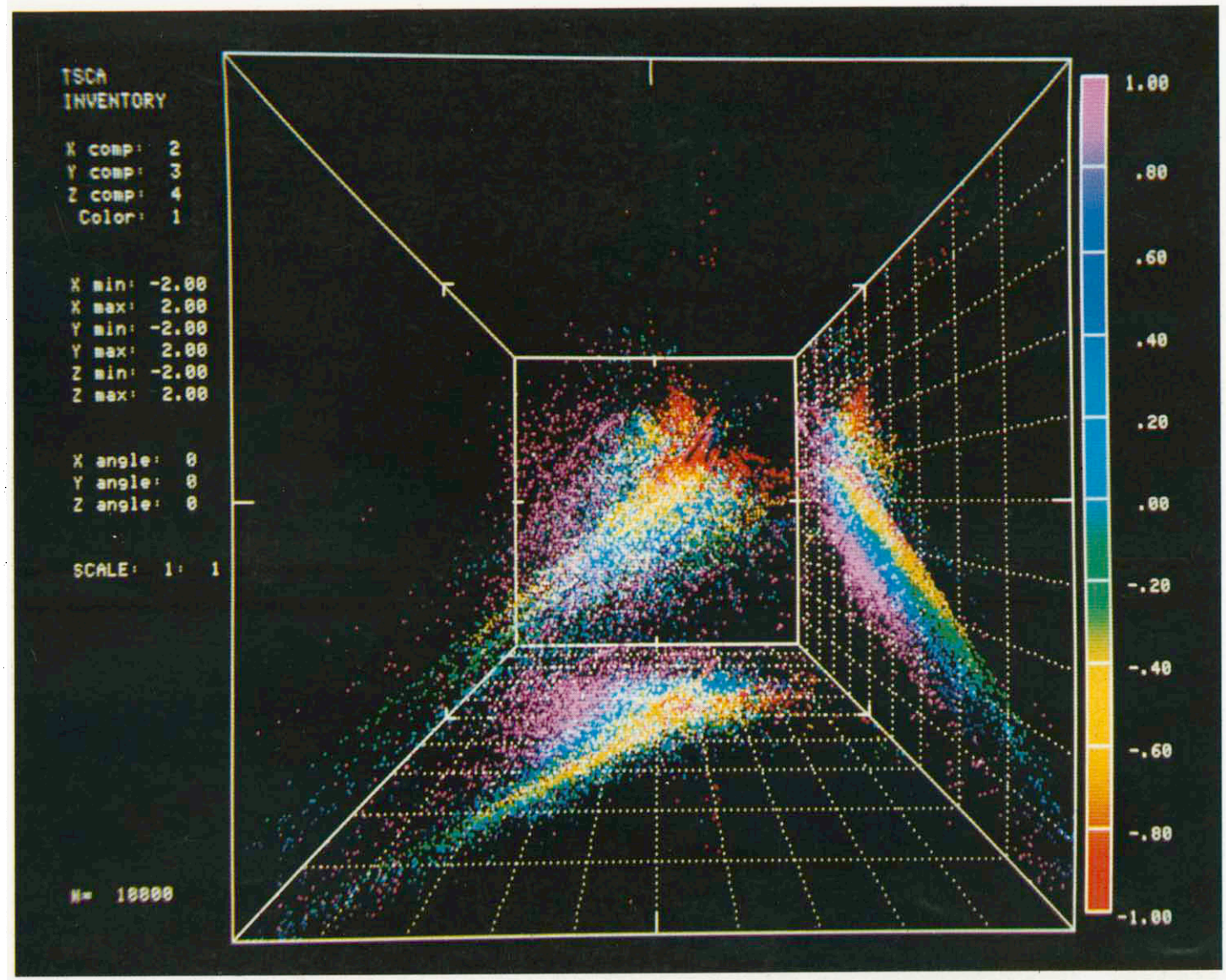

Figure 2. Graphical representation of the chemical universe where the X, Y, Z and color axes are the second, third, fourth and first dimensions, respectively.

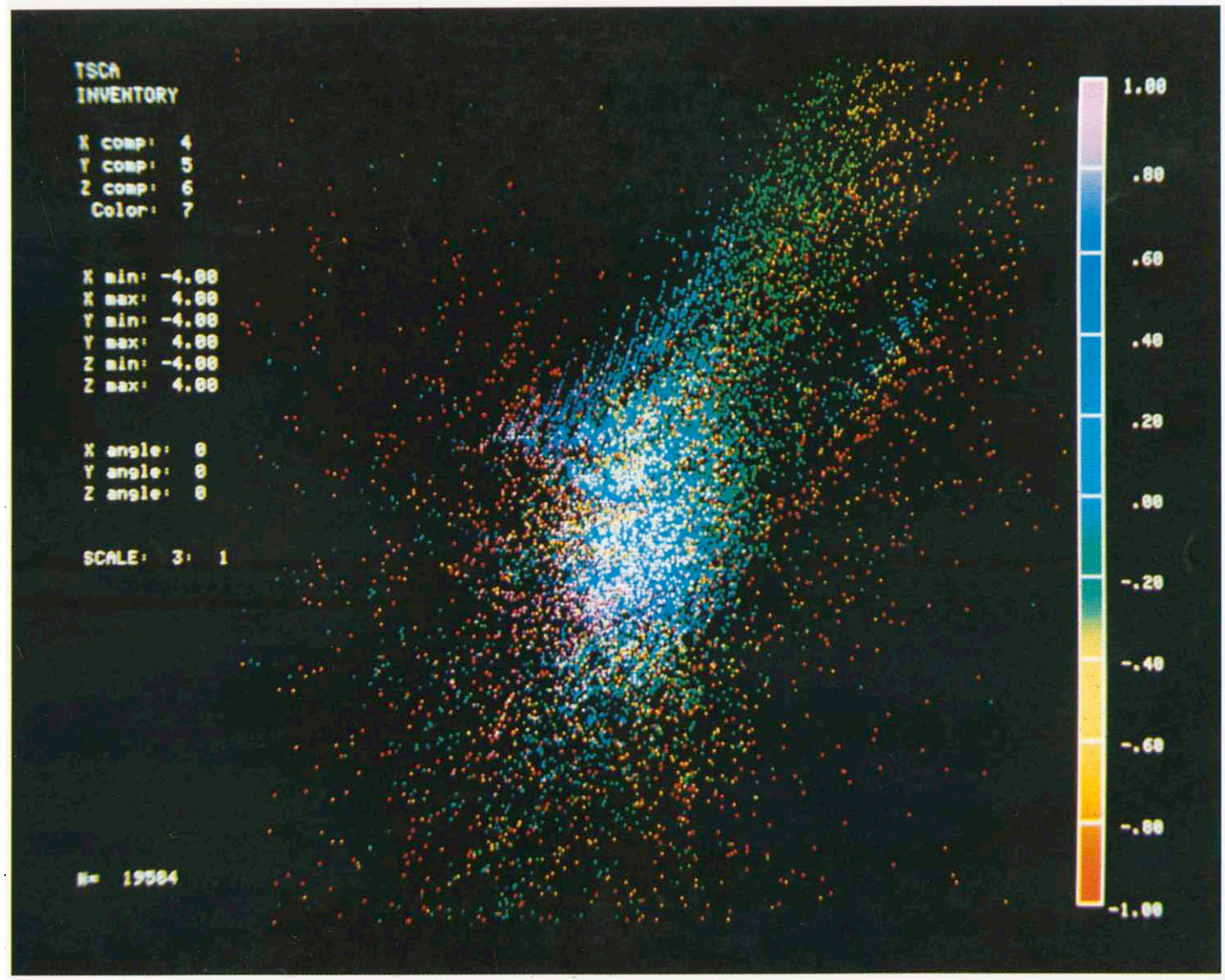

Figure 3. Graphical representation of the fourth, fifth, sixth and seventh dimensions of the chemical universe on the X, Y, Z and color axes, respectively.

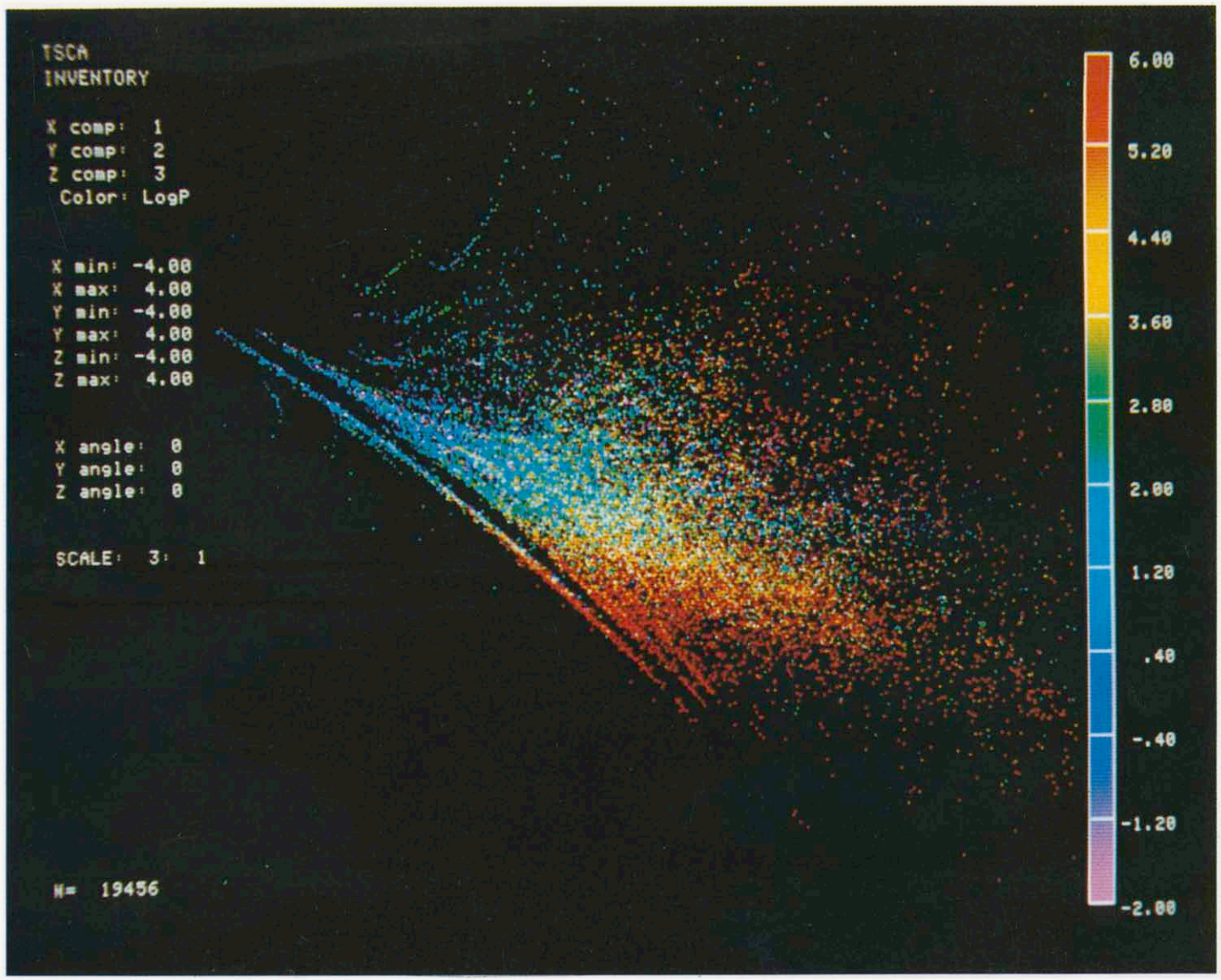

Figure 4. Close up of the first three dimensions of the chemical universe where color is varied (blue to red) proportional to Log P. Blue dots represent water soluble structures and red dots represent extremely lipid soluble structures.

Table 1. Interpretations and examples of extremes for 8 principal components calculated from 90 variables based on connectivity indices for 19,972 industrial chemicals.

Principal Component	Eigen-value	Variation explained %	Low values of principal	High values of principal component
1	47.36	52.6	small molecules	large molecules
2	12.14	13.5	few branches on molecule	multi-branched molecules
3	10.53	11.7	non-cyclic molecules	multi-cyclic molecules
4	5.19	5.8	7th to 8th order cycles	3rd to 4th order cycles
5	3.13	3.5	molecules with single bonds and simple branching patterns	multi-branched molecules with double or triple bonds and/or with many heteroatoms
6	2.83	3.2	complex branching patterns and multi-cyclic molecules with few heteroatoms	complex 3rd and 4th order cyclic molecules
7	1.74	1.9	5th to 7th order cycles	complex valence-corrected branches chemicals with many heteroatroms
8	1.22	1.4	short chain molecules with complex heteroatom branches	long chain molecules with few heteroatoms

In summary, we have developed a stable chemical structure space based on graph theoretic indices and molecular topology. The coordinates for the location of any chemical can be computed from structure. A computer graphics system permits the exploration of the space around the structure, and nearest neighbors appear to be structurally similar to a given chemical. Regions of space are associated with chemical properties such as free-energy; however, substantial work compiling systematic property data bases must be completed before the multi-dimensional space can be tested for predictive power.

REFERENCES

1. L.P Hammett, Physical Organic Chemistry, Second Edition (McGraw-Hill Book Company, New York, 1970), 420 pp.
2. R.F. Gould, Biological Correlations - The Hansch Approach, Advances in Chemistry Series No. 114 (ACS, Washington, D.C., 1972), 340 pp.
3. R.D. Cramer, J. Am. Chem. Soc. 102(6), 1837-1849 (1980).
4. R.D. Cramer, J. Am. Chem. Soc. 102(6), 1849-1959 (1980).
5. R.C. Reid, Fluid Phase Equilibria, 13, 1-14 (1983).
6. W.J. Dunn and S. Wold, Bioorg. Chem. 9, 505-523 (1980).
7. P.E. Long, An Introduction to General Toxicology (Merrill Publ. Co., Columbus, Ohio, 1971), 281 pp.
8. M. Randic, J. Am. Chem. Soc. 97, 6609-6613 (1975).
9. L.B. Kier, and L.H. Hall, Molecular Connectivity in Chemistry and Drug Research (Academic Press, New York 1976), 257 pp.
10. A. Sabljic, and N. Trinajstic, Acta Pharm. Jugosl. 31: 189-214 (1981).
11. N.H. Nie, C.H. Hull, J.G. Jenkins, K. Steinbrenner, and D.H. Bent, Statistical Package for the Social Sciences (McGraw-Hill, New York, 1975), p. 675.
12. G.D. Veith, D.L. DeFoe, and B.V. Bergstedt, J. Fish. Res. Board Can. 36, 1040-1048 (1979).
13. A. Leo, Log P values computed via CLOGP, Pomona College MEDCHEM Project, Claremont, CA.