Stigmata: An Algorithm To Determine Structural Commonalities in Diverse Datasets

N. E. Shemetulskis,[†] D. Weininger,[‡] C. J. Blankley,*,[†] J. J. Yang,[‡] and C. Humblet[†]

Parke-Davis Pharmaceutical Research Division of the Warner-Lambert Company, 2800 Plymouth Road,
Ann Arbor, Michigan 48105, and Daylight Chemical Information Systems, Inc., 419 East Palace Avenue,
Santa Fe, New Mexico 87501

Received October 31, 1995[⊗]

An algorithm, Stigmata, is described, which extracts structural commonalities from chemical datasets. It is discussed using several illustrative examples and a pharmaceutically interesting set of dopamine D_2 agonists. The commonalities are determined using two-dimensional topological chemical descriptions and are incorporated into the key feature of the algorithm, the modal fingerprint. Flexibility is built into the algorithm by means of a user-defined threshold value, which affects the information content of the modal fingerprint. The use of the modal fingerprint as a diversity assessment tool, as a database similarity query, and as a basis for color mapping the determined commonalities back onto the chemical structures is demonstrated.

INTRODUCTION

Chemical structure databases are an invaluable resource for many industries. The increase in the availability of electronic structural information continues to expand the utility of mining such data for decision making purposes.[1] Mining chemical databases requires analysis tools and structural query handling, which, in turn, depend on electronic definitions of chemical structures. Early work in chemical information systems defined structures in terms of chemical fragment codes[2] which later gave way to linear notation schemes such as WLN (Wiswesser line-formula notation)[3] and the more recent SLN (Sybyl line notation)[4] and SMILES (simplified molecular input line entry system).[5] These linear notations define molecular structures by alphanumeric strings. Such definitions can be translated into connection tables or into chemical graphs which are topological molecular descriptions based on two-dimensional connectivity. Database searching algorithms have expanded search queries to include pharmacophore definitions and Euclidean distance measures to find three-dimensional structural hits.[6,7]

With the ever increasing size of chemical databases, chemical information systems continually face the challenge of finding efficient chemical structural representations and similarity searching approaches. A common strategy to increase the efficiency of database searching has been to translate molecular structural representations into binary strings. The structural key approach uses representations of structures which reflect the presence or absence of predefined functional groups. Such definitions are dependent on *a priori* substructural definitions.[8] A fragment dictionary approach to molecular representation uses a training set of molecular structures to generate fragment descriptors which include both topological and generic information.[9] Another approach defines molecular fingerprints using a binary representation of molecular structure generated from the hashing of unique substructural paths.[10] Molecular fingerprints are derived directly from the structures, so the strength of this approach is its independence from an external database of functional definitions and its inclusion of topological structural information. A limitation of any topological molecular description resides in the choice of a maximum path length cutoff. An increase in the cutoff path length decreases the efficiency of the fingerprinting algorithm. Molecular fingerprints are primarily used to efficiently search databases and to analyze chemical similarity. A balance, therefore, between structural information content and the efficiency of fingerprinting databases of thousands of structures determines an appropriate path length cutoff.

The accuracy of chemical information retrieval systems depends on numeric similarity definitions which are chemically meaningful. Similarity assessments of binary molecular representations commonly use the Tanimoto coefficient (T_c) given below to define pairwise molecular similarity.[11]

$$T_c = N(A\&B)/[N(A) + N(B) - N(A\&B)] = N(A\&B)/N(A|B)$$

This coefficient is a ratio between the number of common set bits between two binary strings, A and B, and the number of bits set by both strings. $N(A\&B)$ is the number of common bits. $N(A)$ and $N(B)$ are the number of set bits in compounds A and B, respectively. $N(A|B)$ is the number of bits set by the logical OR, *i.e.*, union, of the two bit strings. This coefficient ranges from zero to one, where zero indicates no common bits and one reflects that all bits are shared between two structures, e.g., the Tanimoto coefficient of a molecule compared with itself is one. Numerical approaches to both chemical structural definitions and accurate retrieval systems using similarity assessments based on such measures still require a bridge from the mathematical descriptions back to the chemical structures.

Applications of chemical database analysis using numerical similarity assessments are of increasing importance in the pharmaceutical industry.[11,12] Corporate databases represent the chemical libraries which are used in high volume screens in the quest for novel leads for drug discovery programs. The search for novelty has prompted recent attempts to diversify corporate libraries using strategies based on chemical database analysis.[13] Diversification strategies are aimed at increasing the probability and frequency of the discovery

* Author to whom correspondence should be addressed.

[†] Parke-Davis Pharmaceutical Research Division.

[‡] Daylight Chemical Information Systems, Inc.

[⊗] Abstract published in *Advance ACS Abstracts,* April 15, 1996.

S0095-2338(95)00169-7 CCC: $12.00

of novel leads, either through database acquisitions or combinatorial chemistry. Many companies which exploit combinatorial chemistry create diverse libraries using some form of library design. These designs depend on the availability of chemical databases in electronic format and the use of computational tools for the analysis.[14−21]

Novel lead structures arise as hits in the high volume screening of chemical libraries. These structures are commonly used as similarity queries on corporate databases or commercially available databases. The rapid identification of structures similar to potent ligands provides an efficient means to develop structure−activity relationships around a particular ligand. An analysis of the hits themselves also provides information for pharmacophore models which in turn can be used as database templates for *de novo* design efforts.

Ligands from a mass screen or structures found from database similarity searches can contain varying degrees of structural diversity. The commonality determined from a set of potent ligands could define features which explain the binding affinity. One challenge posed to any tool for the analysis of diverse datasets is that commonalities may not exist in all structures under consideration. Algorithms which search for common substructures, for example, usually require that the substructures exist in all structures in the dataset. Such algorithms have been widely used in chemical information retrieval systems and retrosynthetic analyses.[22−27]

We describe in this paper a new program, Stigmata, which was developed to find structural commonalities in sets of chemical compounds. Stigmata differs from maximal common subgraph algorithms by incorporating flexibility to allow the detection of commonalities which may exist only in fractions of the dataset. Stigmata also provides a chemically interpretable mapping between chemical similarity measured using topologically based chemical descriptors and chemical structures. Such information can be utilized to query a database, to develop predictive models for novel drug design, or to devise templates for combinatorial library design.

METHODOLOGY

The Stigmata algorithm is divided into an analysis tool and a graphical tool to visualize the numerical results. The numerical tool employs the Daylight Chemical Information Systems, Inc. molecular fingerprint representations of chemical structures. Structures are input using the SMILES notation.[5] The molecular fingerprint is a 2048 bit string in which unique bond paths emanating from each atom in a molecule ranging in length from zero (atom type) to length seven have been hashed resulting in several bits being set throughout the entire bit string for each unique path. From the dataset of molecular fingerprints the key feature of the algorithm is generated. This 2048 bit string is called a modal fingerprint, and it contains the common bits found in the molecular fingerprints in the input dataset.[28,29] The degree to which bits have to be in common in the input dataset in order to set a bit in the modal fingerprint is determined from a user defined threshold value. This value ranges from 50% to 100%. A bit will be set in the modal fingerprint if it is set in at least the threshold percentage of molecules in the input dataset. A threshold value of 50% will extract common bits which are set in at least half of the input dataset. A threshold value of 100% will create a modal fingerprint of bits which are set in all structures in the input dataset. A modal fingerprint does not necessarily represent a complete molecular structure; rather it represents common paths that exist in the structures of the input dataset. It could contain, for example, substructures such as a carbonyl or an amino group, but it also could contain paths or fragments which contain part of a ring such as the path [chlorine-aromatic carbon-aromatic carbon] in *p*-chlorophenol. It is important to keep in mind in the comparisons that follow that the modal fingerprint may represent a single common path or several common paths, or it may in fact be a complete molecular structure. It is defined both by the degree of similarity which exists in the input dataset and the threshold criteria applied by the user.

The back extraction to a functional representation of the modal fingerprint is not a simple task and is not done by the algorithm. Instead a graphical tool provides a visual perspective of the numerical results. The graphical tool colors the atoms of each substructure using one of two color schemes based on a similarity measure, ALAB. ALAB is calculated for each atom in each structure from atom fingerprints and the modal fingerprint. An atom fingerprint is a 2048 bit string comprised of bits which represent all paths containing the atom which contribute to the molecular fingerprint. ALAB is defined as follows:

$$\mathrm{ALAB}_{\mathrm{a}} = N(\mathrm{a\&C})/N(\mathrm{a})$$

N(a&C) is the number of bits in common between the atom fingerprint and the modal fingerprint. N(a) is the number of bits set in the atom fingerprint. ALAB ranges from zero to one, where one indicates that all paths in the molecule containing the atom are part of the modal fingerprint and zero labels the atom as a nonmember of the common paths contained in the modal fingerprint. The ALAB values for each atom in each structure are mapped to one of two color schemes. In the coarse four color scheme, atoms with ALAB values equal to zero are colored red, atoms with ALAB values which fall in the range $0 < \mathrm{ALAB} \leq 0.5$ are colored green, atoms with ALAB values in the range $0.5 < \mathrm{ALAB} < 1$ are colored blue, and atoms with ALAB values equal to one are colored white. For a more detailed examination, a second color scheme which has a ten color gradient following the temperature color scale has been provided. As in the four color scheme, an atom will be colored red if its ALAB value is equal to zero and white if its ALAB value is equal to one. Intermediate values will be colored one of eight colors spanning the temperatures scale spectrum from orange to blue. For either coloring scheme, atoms which appear in red are not part of the common functionality of the input dataset, and atoms which appear in white are part of the common functionality of the input dataset. This provides a convenient mechanism for visualizing the paths in each molecule which are part of the common paths that exist in the dataset. We have found the coarse scheme to be useful when comparing large numbers of structures simultaneously, while the finer scheme may provide more detailed information when only a few compounds are being compared.

The atom coloring is only one mechanism for interpreting the structural similarity. Three numerical similarity descriptors are also calculated to aid in the structural analysis. Stigmata calculates these descriptors using the binary representations for all molecular structures, their atoms, and the commonality between them. The first descriptor MSIM is defined by the following Tanimoto coefficient:

$$\mathrm{MSIM} = N(\mathrm{A\&C})/[N(\mathrm{A}) + N(\mathrm{C}) - N(\mathrm{A\&C})] = N(\mathrm{A\&C})/N(\mathrm{A|C})$$

N(A&C) is the number of bits in common between compound A and the modal fingerprint C. N(A) is the number of bits set in the molecular fingerprint of compound A, and N(C) is the number of bits set in the modal fingerprint. Traditional Tanimoto coefficients measure similarity between molecular fingerprints of chemical structures. In general, the modal fingerprint will not necessarily represent a whole molecule or even a standard substructure. Thus MSIM is a Tanimoto coefficient in which only one of the two items being compared is necessarily a complete molecular structure. This difference needs to be kept in mind when interpreting this number.

The need for a second similarity measure to be used in conjunction with MSIM was prompted by the realization that MSIM could be small either because N(A&C) is small or N(A) is large. Structure A is either missing some of the paths in the modal, or it contains paths which are not part of the modal fingerprint. The second similarity measure, MODP (modal percent), is defined as follows:

$$\mathrm{MODP} = N(\mathrm{A\&C})/N(\mathrm{C})$$

N(A&C) and N(C) are the same as defined above for MSIM. MODP ranges from zero to one and reflects the fraction of bits in the modal fingerprint that are set by compound A. Since $N(\mathrm{A|C}) \geq N(\mathrm{C})$, MODP ≥ MSIM. A molecule with a MODP value of one contains within it all of the common paths determined by the algorithm. A molecule with a MODP of one and an MSIM of one not only contains the modal but also *is* the common structure contained in at least the threshold percentage of compounds in the input data set. A molecule with a MODP value of one and an MSIM value of less than one contains unique paths which are not part of the commonality that exists in the input dataset. One can put these descriptors in the context of molecular comparisons made in database search queries. MODP can be interpreted as a substructural comparison score. It reveals the extent to which a structure contains the common paths in the dataset. The MSIM score can be equated with a similarity score. It reveals the similarity between a structure and the common paths in the dataset.

A third similarity measure, RMINF, describes the relationship between the modal fingerprint and the smallest molecular fingerprint in the training dataset, i.e., the structure with the fewest bits set in its molecular fingerprint. Stigmata calculates RMINF according to the equation below:

$$\mathrm{RMINF} = N(\mathrm{C})/N(\mathrm{S})$$

N(C) is the number of bits set in the modal fingerprint. N(S) is the number of bits set in the smallest molecular fingerprint. N(C) is dependent on the threshold value, and therefore the RMINF value at a given threshold provides a measure of the size of the modal fingerprint relative to a structure in the dataset. For a diverse set of ligands one would expect small RMINF values at large threshold values, since the more stringent the commonality criterion, the smaller the number of common bits that will be found in the dataset. Analysis of RMINF values versus threshold values provides a mechanism for choosing an operational threshold value and for determining the diversity of the dataset.

Stigmata's numerical tool operates on chemical datasets represented in SMILES notation. A SMILES string is a hydrogen suppressed two-dimensional connectivity representation of a molecular structure. A molecular identifier, either a name or registration number, is also required. Stigmata generates the Daylight standard format TDT (Thor Data Tree) file.[10] This file contains fields for the SMILES, molecular name, MSIM, MODP, and atom ALAB information for each structure contained in the input data set. This TDT file can be translated into a visual display using the visualization tool, xvstigmata. Both tools are toolkit programs which employ Daylight Chemical Information Systems, Inc. Toolkits.[30] Stigmata sorts structures first on MODP and then secondarily on MSIM. Structures are ordered such that those which contain the largest percentage of the common features and minimal excess functionality will appear at the top of the output listing. The first 64 structures are displayed in an XView window. MSIM and MODP values appear in a label field above each structure. The ALAB values for each atom in each structure are mapped to one of the two color schemes.

There are three command line options to the program. The first of these generates both a TDT file and an ASCII output file containing the name, MSIM and MODP. The threshold value appears at the end of the file along with the name of the structure with the most bits set in its molecular fingerprint and the name of the structure with the fewest bits set in its molecular fingerprint. The fraction of bits set in the modal fingerprint relative to either the largest or the smallest fingerprints in the input dataset is given. This file provides a mechanism for further analysis of the Stigmata similarity measures with external routines.

A second command line option is a modal export option. The modal fingerprint determined from an input dataset can be compared to a second specified dataset. The TDT output file contains the similarity comparisons between the compounds of the second dataset and the modal fingerprint from the first input dataset. This option is useful if the second dataset is small and visualization with the xvstigmata tool is desired.

The third command line option is an extension of the modal export option to search databases containing thousands of structures. This option provides similarity searching of a database using the modal fingerprint as the query. When this option is specified, the modal fingerprint is determined from the input dataset, and MODP and MSIM values are generated for a specified database which is in the same format as the input dataset. Rather than a TDT file, an ASCII file is generated which contains the name and MODP and MSIM values for all compounds in the comparison database which have MODP values greater than 50%.

APPLICATIONS

Bromophenol and Substituted Indoles Dataset. To illustrate how the program works, a simple dataset given in Figure 1a consisting of substituted benzene rings was analyzed with Stigmata using a threshold of 1.0 and a

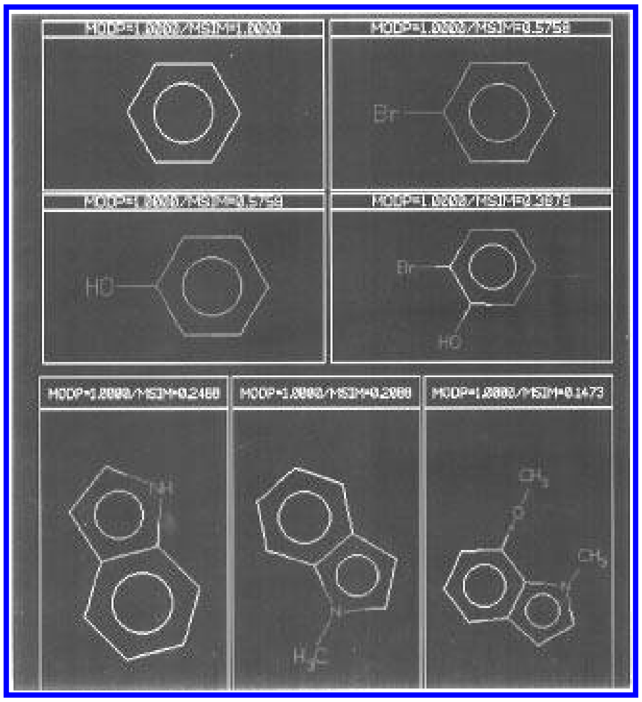

Figure 1. A visual display of the Stigmata analysis of benzene, phenol, bromobenzene, and *o*-bromophenol at a threshold value of 1.0 is given in part a (top). The modal fingerprint was exported from the dataset in part a and used in a Stigmata analysis of the three substituted indoles displayed in part b (bottom). The visualization tool employed the ten color scheme to display the ALAB atom information.

threshold value of 0.5. The results are given in Figures 1a and 2a, respectively. The algorithm was run a second time at each threshold value with the modal export option set. The modal found in the substituted phenols was compared to a second dataset containing substituted indoles. Figures 1b and 2b provide a display of these results. The fine coloring scheme was used in Figures 1 and 2.

Piperazine Series. To demonstrate further capabilities of the algorithm an example dataset of 148 structures containing piperazine was created. The structures were obtained from a substructure search on the Maybridge database using piperazine as the query. The Maybridge database is a subset of 41 912 structures from the Available Chemicals Directory (ACD).[31] We found 147 structures and these, combined with piperazine, were used as input to Stigmata. A UNIX shell script was written which generated TDT files for the piperazine dataset analyzed at threshold values ranging from 0.5 to 1.0, incremented by 0.1. The shell script also generated an ASCII file containing threshold, RMINF, and the name of the compound with the smallest fingerprint, which in this case was piperazine. A plot of this data is given in Figure 3a. The first structure in the output file obtained for each run at a different threshold value is displayed in Figure 3b using the coarse coloring scheme.

Dopamine Series. A set of 33 high affinity dopamine D_2 ligands[32] was analyzed with Stigmata using the UNIX shell script described above for the piperazine study. A plot of RMINF versus threshold value is given in Figure 5a. Dopamine has the smallest fingerprint for this dataset. The first structure in the output file which was obtained for each threshold value is displayed in Figure 5b. Figure 4 displays the full output file from a Stigmata analysis run at a threshold of 0.5. Since the MODP and MSIM values are illegible in this figure, they are listed in Table 1 with the compounds ordered as they appear in the display. In Figures 4 and 5 the coarse coloring scheme was used for the color depictions.

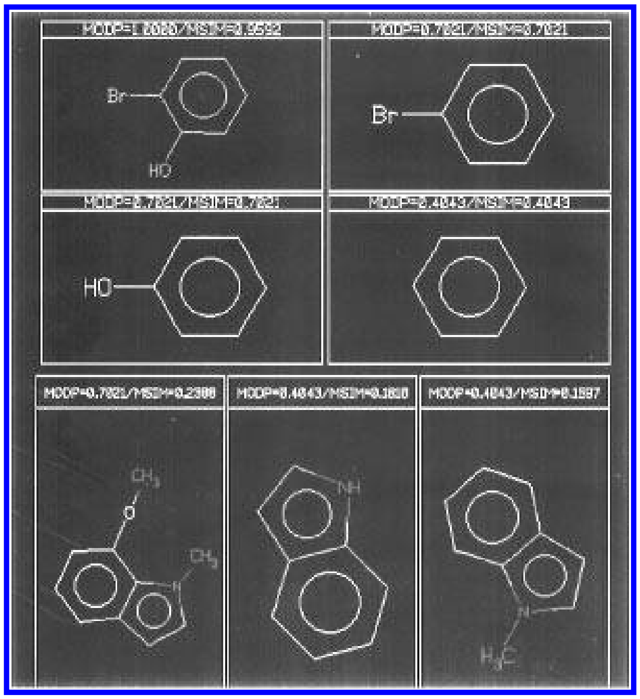

Figure 2. The same datasets which were analyzed in Figure 1 are displayed in this figure in which the Stigmata analysis was run at a threshold of 0.5. The ten color scheme has been used to display the ALAB atom information.

The modal obtained from the dopamine dataset at a threshold of 1.0 was used as a query to search the Maybridge database. For this analysis the database was subdivided into three subgroups of 10 000 structures and one subgroup of 11 912 structures to enable the algorithm to run in parallel on the entire database of 41 912 structures. The database search option has the most demanding requirements for CPU cycles. The algorithm required half an hour of CPU time using four processors of a Silicon Graphics ONYX workstation. The structures with the highest similarity to the modal fingerprint are given in Figure 6a, and the MODP and MSIM values are given in Table 2. The same searching procedure was used in Figure 6b for a modal fingerprint determined from a threshold of 0.5, and the corresponding MODP and MSIM values are given in Table 3.

DISCUSSION

Bromophenol and Substituted Indoles Dataset. In order to appreciate the information contained in the similarity measures and the algorithmic capabilities of Stigmata, it is instructive to first analyze a simple dataset. In Figure 1a the results from a Stigmata analysis of four structures, benzene, phenol, bromobenzene, and *o*-bromophenol are displayed. The threshold was set to 1.0, and, therefore, common paths which are part of the modal fingerprint appear in all of the structures. With a threshold value of 1.0, all structures have a MODP value of 1.0. The MODP of a molecule is an upper bound for its MSIM value and the difference between a molecule's MSIM value and its MODP value increases as the number of unique paths in the molecular structure increases. For this dataset, benzene is present in all four structures and therefore, benzene appears

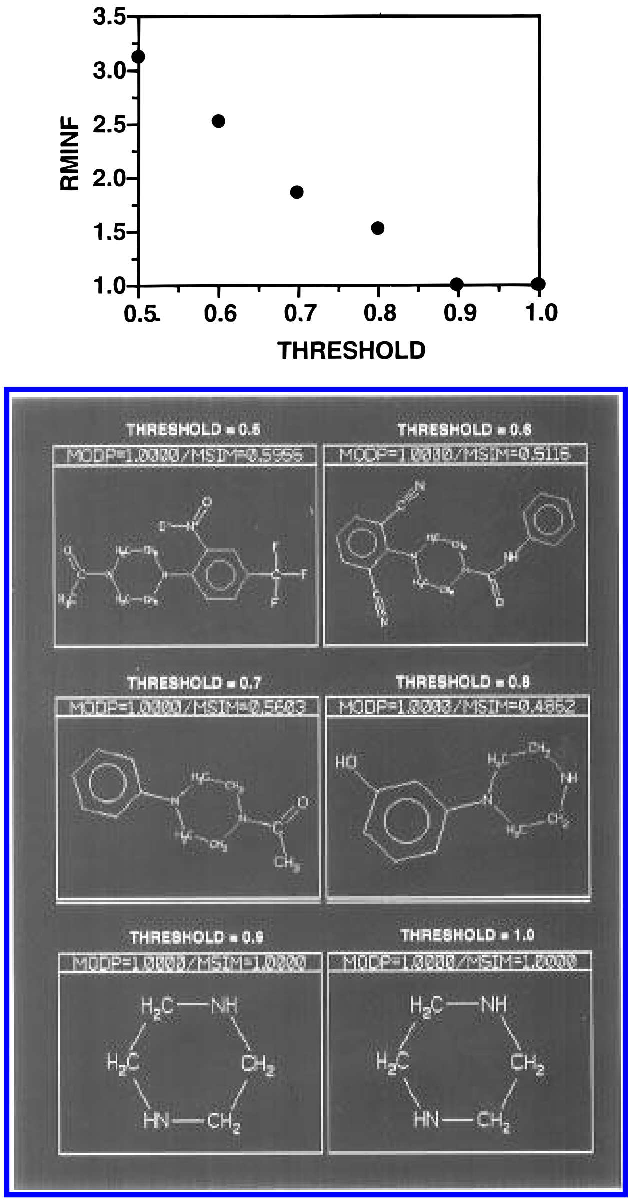

Figure 3. A plot of RMINF determined from a dataset of structures found in a substructural search of the Maybridge database for piperazine are plotted versus threshold value in part a (top). The first structure in the output file from each run at thresholds from 0.5 to 1.0 are given in part b (bottom), using the four color scheme. The number of bits set in the molecular fingerprint of piperazine constitute the denominator of RMINF.

as the first compound in the output file, with MODP and MSIM values of 1.0. Benzene not only contains the commonality of the dataset, but it is also the structural representation of the modal fingerprint. The MSIM values for the substituted ring systems in Figure 1a deviate from 1.0 the more highly substituted the structure. The features which are unique to a given structure and are therefore not part of the modal fingerprint appear in red. The color of the aromatic ring in all cases except for benzene is blue, since the carbon atom scores are slightly less than one due to the paths generated from the ring substitutions. All carbons in benzene are white, since all paths through each carbon atom are part of the modal fingerprint.

The modal fingerprint was exported from the substituted aromatic dataset and compared to a set of substituted indoles: indole, 1-methylindole, and 1-methyl-7-methoxyindole. The results are displayed in Figure 1b. The MODP values are all 1.0, reflecting the fact that benzene is an embedded substructure in indole. The bright orange color of the phenyl substructure of the indole ring highlights the low similarity to the modal fingerprint. The burnt orange color of the pyrrole ring and the red methoxy and methyl substitutions correlate with the decreasing MSIM values. These reflect the greater numbers of unique paths in the substituted indoles which result in the decreasing similarity to benzene.

Table 1. Stigmata Results for D2 Agonist Dataset at Threshold = 0.5 (Compounds Numbered as Illustrated in Figure 4)

compd no.	MODP	MSIM
1	1.0000	0.5992
2	0.9931	0.6128
3	0.9862	0.6111
4	0.9724	0.5779
5	0.9724	0.4764
6	0.9655	0.7568
7	0.9586	0.5206
8	0.9517	0.7797
9	0.9310	0.7627
10	0.9310	0.7337
11	0.8966	0.7471
12	0.8483	0.4522
13	0.8276	0.4412
14	0.8207	0.6723
15	0.8138	0.4419
16	0.8138	0.4403
17	0.8069	0.6882
18	0.8069	0.4366
19	0.8000	0.6339
20	0.8000	0.5743
21	0.8000	0.5659
22	0.7793	0.6457
23	0.7724	0.6222
24	0.7172	0.2413
25	0.6966	0.2701
26	0.6897	0.2762
27	0.6828	0.6828
28	0.6828	0.3204
29	0.6828	0.2690
30	0.6414	0.6327
31	0.6069	0.6069
32	0.6000	0.6000
33	0.5517	0.5517

Stigmata's flexibility to search for commonalities which exist in subsets of a dataset is examined in Figure 2a. In Figure 2a the results are displayed from a Stigmata analysis on the same dataset in Figure 1 using a threshold value of 0.5. Features appear in the modal fingerprint if they appear in at least two of the four structures. With a threshold of 50%, the first structure which appears in Figure 2a is *o*-bromophenol. Two structures have hydroxy substitutions and two have bromine substitutions. *o*-Bromophenol therefore has a MODP value of 1.0 and the highest MSIM value. Its MSIM value is less than 1.0, since the modal fingerprint does not contain paths for a disubstituted system. The modal fingerprint was compared against the substituted indole structures, and the results are displayed in Figure 2b. None of the structures contain all of the features in the modal fingerprint, as shown by MODP values all less than 1.0. The structure with the highest MODP value is 1-methyl-2-methoxyindole, since it contains the benzene ring and the oxygen of the methoxy and the substitution shares some common paths with the hydroxyl containing paths in the modal fingerprint. The bright orange color of the methoxy oxygen compared to the red methyl substitution is consistent

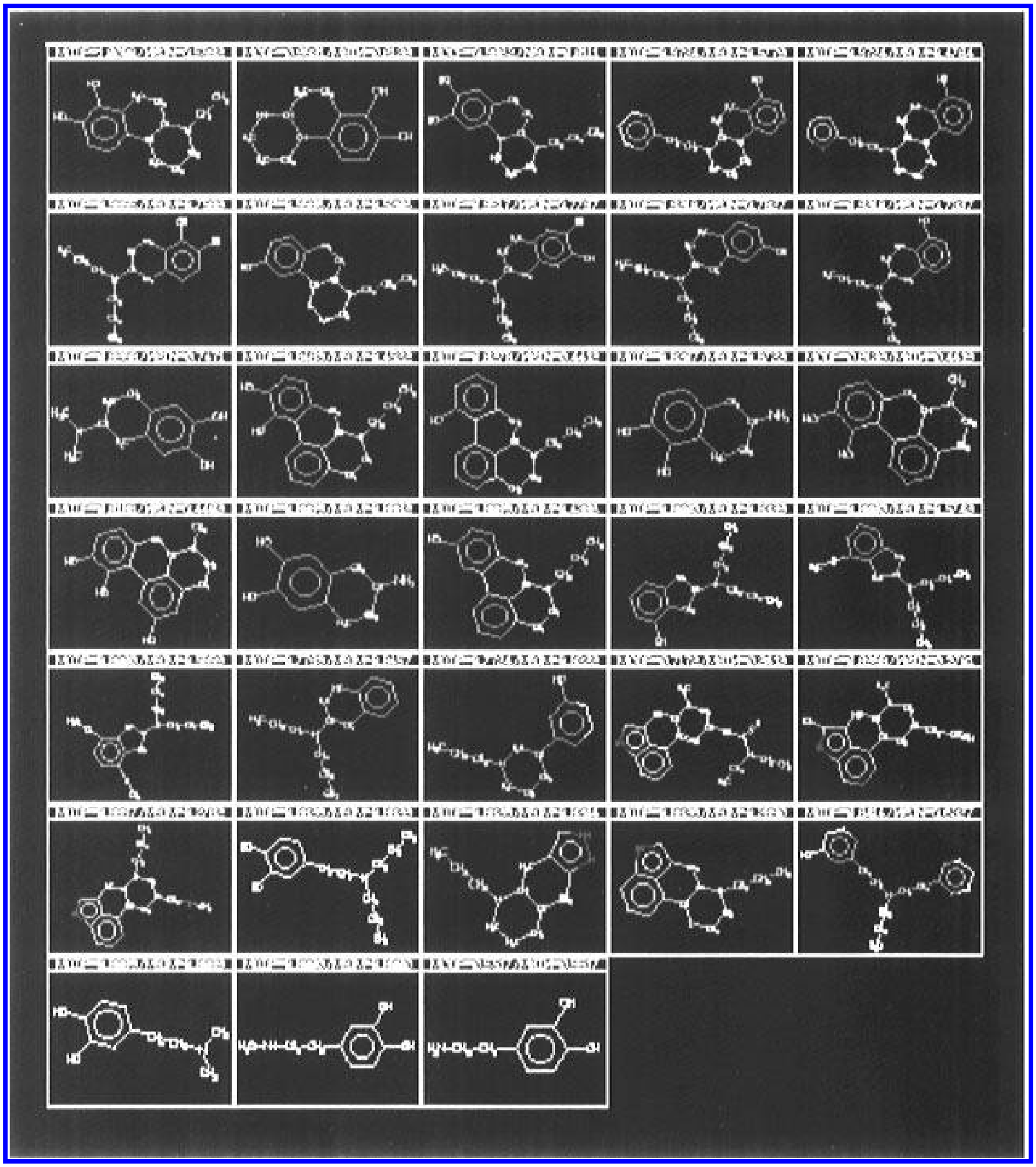

Figure 4. The Stigmata analysis of a dataset of high affinity dopamine D_2 agonists at a threshold of 0.5 are displayed in this figure. The four color scheme is used in the representation and the MODP and MSIM values for each compound reading left to right are given in Table 1.

with the higher MSIM value of 1-methylmethoxyindole relative to indole and 1-methylindole.

Piperazine Dataset. Before considering the analysis of chemical structures containing piperazine, it is important to discuss first the general relationship between the threshold value and the number of paths in the modal fingerprint. Certainly commonalities among chemical structures are dependent on the chemical makeup of the structures themselves. The user controlled threshold value provides a lower bound on the number of structures which must contain a common path if that path is to be part of the modal fingerprint. At a low threshold value of 50% the modal fingerprint contains all paths which are present in at least half of the dataset. The number of bits set in the modal fingerprint at this threshold value is at its maximum and will either remain constant or decrease as the threshold value is increased. In other words, the higher the threshold value the more stringent the criterion on the creation of the modal fingerprint. At a threshold value of 1.0, the number of bits set in the modal fingerprint is bounded by the number of bits set in the smallest molecular fingerprint, since paths at this threshold level must exist in all structures to appear in the modal fingerprint. The smallest molecular fingerprint is the fingerprint containing the least number of set bits. Stigmata determines the number of bits in the smallest molecular fingerprint and the number of bits set in the modal fingerprint. RMINF provides an estimate of the amount of information contained in the modal fingerprint relative to

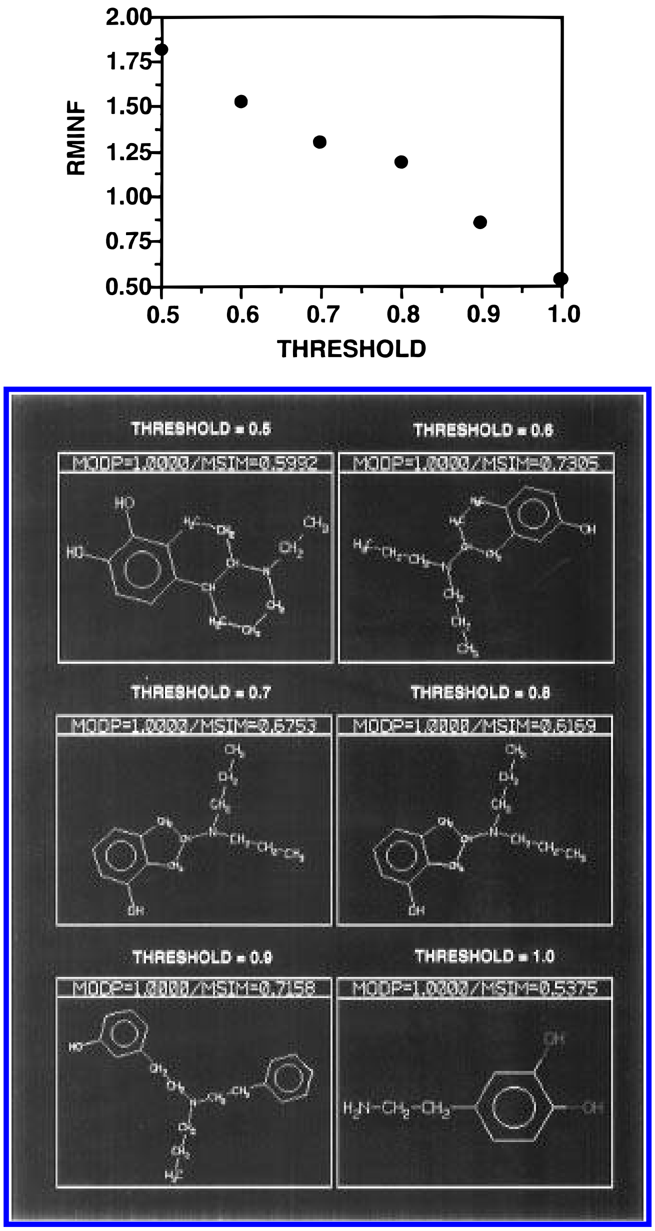

Figure 5. RMINF as determined from the D_2 dataset is plotted versus threshold value in part a (top). The first structure in the output file from each run at threshold values from 0.5 to 1.0 are given in part b (bottom) using the four color scheme. Dopamine is the structure with the least number of bits set in its molecular fingerprint.

that in the smallest molecular fingerprint in the training set. It is important to note that there is not a one-to-one correspondence between set bits and unique paths, since the hashing algorithm results in several bits being set for each unique path and the number of bits is dependent on the path length. A plot of RMINF versus threshold provides a diversity assessment of a chemical dataset. RMINF values of less than 1.0 imply that the commonalities which have been detected are fewer than the unique number of paths in the smallest molecular fingerprint. RMINF values of greater than 1.0 imply that the commonality in the dataset is larger than the number of unique paths in the smallest molecular fingerprint.

A threshold analysis was used to assess the modal information contained in a hitlist of structures from the Maybridge database resulting from a substructural search for piperazine. In Figure 3a, RMINF is plotted versus threshold value. The structure with the smallest fingerprint for this set of 149 structures is piperazine. The threshold values range from 0.5 to 1.0, in increments of 0.1. At a threshold value of 0.5, the modal fingerprint contains more than three times the number of bits set in piperazine, and, at threshold values of 0.9 and 1.0, the bits set in the modal fingerprint are identical to the bits set in the molecular fingerprint of piperazine. In Figure 3b, the first structure in the Stigmata output file at each threshold value is displayed. For threshold values of 0.9 and 1.0, piperazine appears in both cases as the structure with the highest MODP and MSIM value. In both cases MODP and MSIM are 1.0, revealing that the modal fingerprint represents, as expected, the unique paths contained in piperazine. White atom colors reflects that all atoms and paths through the atoms are part of the modal fingerprint. Piperazine is the only commonality existing in 90% of the data from the substructural search. The findings at threshold values less than 90% reveal something previously unknown about the data. At least 80% of the structures from the piperazine substructural search contain nearly all of the paths in *N*-phenylpiperazine. The first structure in the output file is *N*-(3-hydroxyphenyl)piperazine, and it has an MODP value of 1.0 and an MSIM value of 0.4862. The small MSIM value coupled with the green color of the phenol substituent reveals that some paths in this part of the structure are not part of the commonality in the dataset. A scan of the first 64 structures in the output file reveals that Stigmata is sensitive to the fact that less than 80% of the structures contain a phenyl ring. Rather, 80% of the structures contain either a phenyl or a pyridine ring. The modal fingerprint at 80% includes only a portion of the phenyl ring, due to the fact that pyridine contains nitrogen. Once the threshold is decreased to 60%, one finds that at least 60% of the structures contain the unique paths of *N*-phenylpiperazine.

The threshold analysis applied to the piperazine dataset has provided a means of assessing the structural commonalities in a chemical dataset for which some structural information was known. Data coming from a substructural search is certainly less diverse than one coming from ligands found from a binding assay. This leads to the application of a threshold analysis for a set of dopamine ligands.

Dopamine D_2 Ligand Analysis. Stigmata was written as a tool to find structural commonalities in diverse datasets. Typically, diverse datasets of ligands from a binding assay will not necessarily contain similar structural features as well defined as the piperazine or *N*-phenylpiperazine extracted from the previous dataset. Figure 4 displays a dataset of dopamine D_2 high affinity agonists in which a Stigmata analysis has been performed at a threshold value of 0.5. The MODP and MSIM values are given in Table 1. The compounds are numbered by rows reading left to right. Several interesting structures displayed in Figure 4 stand out, since they appear all in white in the structural depictions. The white structures have MSIM values equivalent to their MODP values. A MODP value of less than 1.0 indicates that a structure is missing some of the paths contained in the modal fingerprint. For example, at least 50% of the structures contain all of the parts of *N*,*N*-dipropyldopamine (compound **26**) as a substructure, but its MODP value is only 0.6828.

In Figure 5 a threshold analysis has been performed on the dopamine D_2 dataset for the same range and increment of threshold values as in Figure 3. The ordinate values in

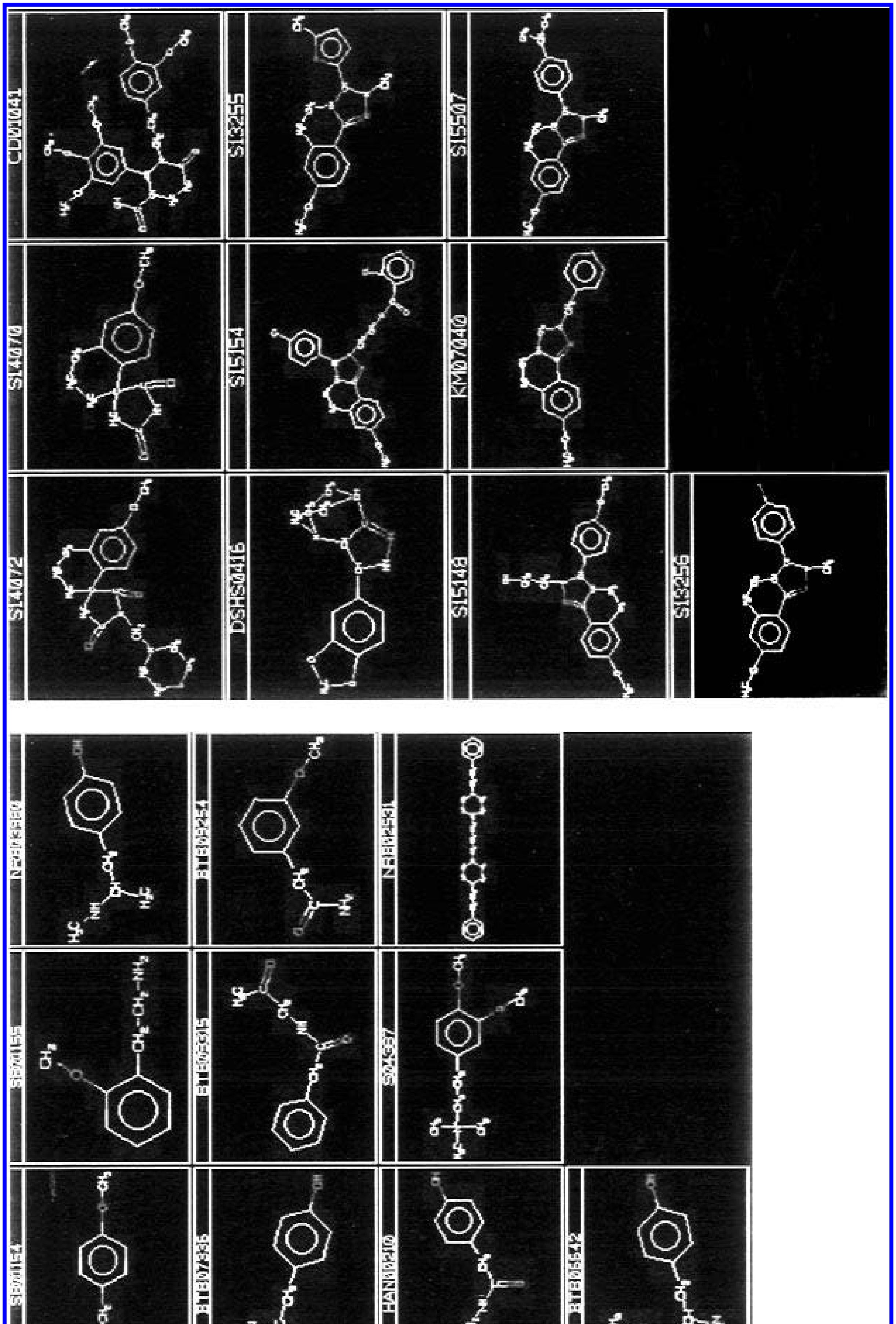

Figure 6. The top ten structures resulting from a database search using the modal fingerprint from an analysis of the D_2 dataset at a threshold of 1.0 are given in part a (left). The MODP and MSIM values can be found in Table 2. The top ten structures resulting from the same dataset dataset run at a threshold of 0.5 are given in part b (right). The MODP and MSIM values are listed in Table 3. The ten color scheme has been used in the color depictions.

Table 2. Top Ten Hits from Search of Maybridge Using the Modal Fingerprint from the D2 Dataset with Threshold = 1.0

Maybridge no.	MODP	MSIM
SB01154*	1.0000	0.5181
SB01155*	1.0000	0.5059
NRB03980*	1.0000	0.4886
BTB07336	1.0000	0.4388
BTB09315	1.0000	0.4343
BTB09254*	1.0000	0.4300
HAN00210	1.0000	0.4135
SO4387*	1.0000	0.4135
NRB02531	1.0000	0.3909
BTB06642	1.0000	0.3874

Table 3. Top Ten Hits from Search of Maybridge Using the Modal Fingerprint from the D2 Dataset with Threshold = 0.5.

Maybridge no.	MODP	MSIM
S14072	0.9310	0.3835
S14070*	0.9172	0.4683
CD01041	0.8483	0.4155
DSHS0416	0.8483	0.3245
S15154	0.8483	0.2668
S13255	0.8414	0.2687
S15148	0.8345	0.3033
KM07040	0.8207	0.3315
S15507	0.8207	0.3190
S13253	0.8207	0.3051

Figure 5a are the RMINF values relative to dopamine which was determined to have the smallest molecular fingerprint in this dataset. The plot in Figure 5a immediately reveals that the commonality in greater than 80% of the structures is less than the number of unique paths in dopamine itself. This dataset as expected is more structurally diverse than the piperazine example. At a threshold value of 1.0, Figure 5b displays dopamine as the structure with the highest similarity to the modal fingerprint. Dopamine's MODP value is 1.0 indicating that the bits set in its modal fingerprint are a subset of the bits set in dopamine's molecular fingerprint. Dopamine's MSIM value and the RMINF value in Figure 5a are equal. The following proof illustrates this relationship. If MODP = 1, then $N(S\&C) = N(C)$. $N(S|C) = N(S|(S\&C)) = N(S)$. Thus, MSIM = $N(S\&C)/N(S|C) = N(C)/N(S)$ = RMINF.

Atom coloring based on atom ALAB values in Figure 5b eliminates paths containing the hydroxyl groups as part of the modal fingerprint, since a red color represents an ALAB value of zero. These groups provide the explanation for the difference between dopamine's MSIM value of 0.5375 and MODP value of 1.0. The first structure in the Stigmata output file in a analysis at a threshold value of 50% has a MODP value of 1.0 and an MSIM value equal to 0.5992. The blue color in the structure now reveals several common paths containing the hydroxyl substitutions on the phenyl ring and the substituted amino group two carbons removed from the phenyl ring to be part of the modal fingerprint.

The modal fingerprint derived from the D_2 agonists at various threshold levels is not a complete molecular substructure. Searching a chemical database for structures similar to the modal fingerprint is a logical next step for chemically available compounds which might also display D_2 agonist activity or which could be used as starting materials for the synthesis of other dopamine D_2 agonists.

Dopamine D_2 Maybridge Database Search. The modal fingerprint generated from a Stigmata analysis at a threshold value of 1.0 was used as a similarity search query on the Maybridge database. Maybridge was selected, since it is a database of chemically available structures and because it was readily available in the proper format for Stigmata. If a set of ligands bind to a particular receptor due to common features which exist in the dataset, then other compounds containing the same features are also worth searching for in available chemical databases. We found 4664 structures with MODP values greater than 0.9. The top ten structures are displayed in Figure 6a, and their similarity scores are given in Table 2. The fine color scheme has been used in the structural depictions. The structure with the highest MSIM value from a Stigmata analysis using a threshold of 1.0 is dopamine. For comparison, dopamine was used as a structural query in a similarity search of the Maybridge database. We found 1317 structures with Tanimoto similarities to dopamine ≥50%. From the ten structures with the highest similarity to dopamine, only five match those displayed in Figure 6a. The matching structures are indicated by an asterisk in Table 2.

The modal fingerprint for the dopamine D_2 agonist dataset at a threshold of 0.5 was then used as a comparison query to the Maybridge database. We found 4816 structures with MODP values greater than 0.5. The top ten structures with the highest MODP values are displayed in Figure 6b using the fine color scheme. The MODP and MSIM values are given in Table 3. The MODP values range from 0.8207 to 0.9310, which reveals that none of the structures contains all of the commonality of the modal fingerprint. The MSIM values range from 0.2668 to 0.4693, which reveals that the structures in Figure 6a contain many paths dissimilar to those which comprise the modal fingerprint. The first structure depicted in Figure 5b having a MODP of 1.0 and an MSIM of 0.5992 was used to search the Maybridge database. This structure contains the commonality of the D_2 agonists but also some unique structural features since its MSIM value is less than its MODP value. We found 553 structures to be similar to the structural query with a Tanimoto similarity greater than or equal to 0.5. Only one of the top ten structures in the similarity search matches the structures found in Figure 6b using the modal fingerprint as the query. This structure is indicated by an asterisk in Table 3. Thus, the modal fingerprint used as a database query has extracted from the Maybridge database in a new and informative way structures that were not found using a similarity search on the most representative structure of the dataset. Furthermore, if it were desired to use one of these compounds as a template for the synthesis of novel D2 candidates, the atom coloring indicates the regions which are most similar and most dissimilar to the modal and thus where synthetic modification should be targeted. It should also be clear that newly conceived, but not yet synthesized, compounds could be evaluated for their similarity to the modal fingerprint and prioritized on this basis.

CONCLUSION

Stigmata is an algorithm which was developed to find commonalities in chemically diverse datasets. The key features of Stigmata, the modal fingerprint determination and the color coded mapping of the commonality information back onto the molecular structures, enable commonalities to be determined and abstract numeric results to be displayed in a form which is interpretable in the structural language of synthetic chemists. An unexpected utility of the algorithm has been implementation of the modal fingerprint directly as a database query, thus providing a novel template for similarity searching. We have demonstrated that potentially interesting compounds in a commercial database can be extracted using this “fuzzy” query. Structures which are part of the hitlist from a modal search were not all found from direct similarity searching of the chemical structure closest in representation to the features contained in the modal fingerprint.

The foundation of Stigmata is the Daylight Chemical Information Systems, Inc. molecular fingerprint. This molecular description has the advantage of being derived directly from the molecular structure and incorporates topological connectivity information. It does not, however, contain three-dimensional structural information. The question arises as to whether the algorithm is transferable to other representations of molecular structure. Other fingerprinting algorithms exist for both two- and three-dimensional chemical structural information.[4,8,33] The extraction of a modal fingerprint from other binary molecular representations is a straightforward extension of the algorithm. It would be less straightforward, however, to adapt the second key feature, the color mapping, to a different fingerprinting algorithm. This is due to the fact that the atom scoring determination is intimately connected to the Daylight fingerprinting algorithm. A fingerprint dependent atom scoring function would have to be developed to enable a sensible and interpretable color mapping using other fingerprinting algorithms.

We are currently exploring practical applications of Stigmata to a variety of chemical datasets to determine its applicability and utility in different drug design scenarios. Among these are template extraction and diversity analysis, which have applications in combinatorial chemistry library design, and topological pharmacophore extraction. Results of these studies will be presented in subsequent communications.

ACKNOWLEDGMENT

N.E.S. would like to thank Jack Delaney for his help in learning both C and Daylight Toolkit programming and also both Jack Delaney and Craig James for initial discussions about the algorithm. We would like to thank John Bradshaw for sharing his early interest in the area of modal fingerprints and Eric Martin for helpful suggestions about the similarity definitions and formulations.

REFERENCES AND NOTES

(1) Maggiora, G. M.; Johnson, M. A. *Introduction to Similarity in Chemistry*; John Wiley & Sons, Inc.: New York, 1990.

(2) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press Ltd.: Letchworth, 1987.

(3) Smith, E. G.; Baker, P. A. *The Wiswesser Line-Formula Chemical Notation (WLN)*; Smith, E. G., Baker, P. A., Eds.; Chemical Information Management Inc.: NJ, 1975.

(4) *SLN Manual*; Tripos Associates, Inc.: St. Louis, MO, 1995.

(5) Weininger, D. SMILES: a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(6) Humblet, C.; Dunbar, J. B., Jr. 3D Database Searching and Docking Strategies. *Ann. Repts. Med. Chem.* **1993**, *27*, 275−284.

(7) Willett, P. *Three-dimensional Chemical Structure Handling*; Research Studies Press: Taunton, 1991.

(8) *MACCS-II*; MDL Ltd.: San Leandro, CA, 1992.

(9) Barnard, J. M.; Downs, G. M. *Fingerprint Descriptor Package*, 3.1; Barnard Chemical Information Ltd.: Sheffield, 1995.

(10) James, C. A.; Weininger, D. *Daylight Theory Manual*; Daylight Chemical Information Systems, Inc.: 1995.

(11) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nonhierarchic Cluster Analysis Methods in Chemical Information Systems: Selection of Compounds fro Biological Testing and Clustering of Substructure Search Output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109−118.

(12) Downs, G. M.; Willett, P. Clustering of chemical structure databases for compound selection. In *Advanced Computer-Assisted Techniques in Drug Discovery;* van de Waterbeemd, H., Ed.; VCH: Weinheim, 1994; pp 111−130.

(13) Shemetulskis, N. E.; Dunbar, J. J.; Dunbar, B.; Moreland, D. W.; Humblet, C. Enhancing the Diversity of a Corporate Database Using Chemical Database Clustering and Analysis. *J. Comput-Aided Molecular Design* **1995**, *9*, 407−416.

(14) Wipke, W. T.; Koehler, R. Characterizing molecular diversity of molecular populations. National Meeting of the American Chemical Society, Anaheim, CA; American Chemical Society: Washington, DC, April 2−6, 1995; Abstract 209.

(15) Siani, M. A.; Weininger, D.; Blaney, J. M.. CHORTLES: A Method for Representing and Searching Oligomeric and Template-based Mixtures. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1026−1033

(16) Siani, M. A.; Weininger, D.; Blaney, J. M.. CHUCKLES: A method for representing and searching peptide and peptoid sequences on both monomer and atomic levels. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 588−593.

(17) Sheridan, R. P.; Kearsley, S. K. Using a Genetic Algorithm to Suggest Combinatorial Libraries. *J. Chem. Inf. Comput Sci.* **1995**, *35*, 310−320.

(18) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(19) *Project Library;* MDL, Ltd.: San Leandro, CA, 1995.

(20) *Legion and Selector*; Tripos Associates: St. Louis, MO, 1995.

(21) Lauri, G.; Bartlett, P. A. CAVEAT: A program to facilitate the design of organic molecules. *J. Comput.-Aided Molecular Design* **1994**, *8*, 51−56.

(22) Brown, R. D.; Downs, G. M.; Willett, P.; Cook, A. P. F. A Hyperstructure Model for Chemical Structure Handling: Generation and Atom-by-Atom Searching of Hyperstructures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 522−531.

(23) Hagadone, T. R. Molecular Substructure Similarity Searching: Efficient Retrieval in Two-Dimensional Structure Databases. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 515−521.

(24) Varkony, T. H.; Shiloach, Y.; Smith, D. Computer-Assisted Examination of Chemical Compounds for Structural Similarities. *J. Chem. Inf. Comput. Sci.* **1978**, *19*, 104−111.

(25) Takahashi, Y. Identification of structural similarity of organic molecules. *Top. Curr. Chem.* **1995**, *174*, 105−33.

(26) Bayada, D. M.; Simpson, A. W.; Johnson, A. P.; Laurence, O. C. An Algorithm for the Multiple Common Subgraph Problem. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 680−685.

(27) Xu, J. Graph Theory: Match, Subgraph Match, Partial Match or Fuzzy Match, personal communication.

(28) We are indebted to J. Bradshaw for informing us of some early formulations of this concept.

(29) Liston, J.; Weibe, W.; Colwell, R. R. Quantitative Approaches to the Study of Bacterial Species. *J. Bacteriol.* **1963**, *85*, 1061−1070.

(30) *Toolkits*, 4.41 ed.; Daylight Chemical Information Systems, Inc.: Irvine, CA, 1995.

(31) *Maybridge database*, 1994 ed.; Daylight Chemical Information Systems, Inc.: Irvine, CA, 1994.

(32) Seeman, P.; Watanabe, M.; Grigoriadis, D.; Tedesco, J. L.; George, S. R.; Svensson, U.; Lars, J.; Nilsson, G.; Neumeyer, J. L. Dopamine D_2 Receptor Binding Sites for Agonists: A Tetrahedral Model. *Molecular Pharmacology* **1985**, *28*, 391−399.

(33) Brown, R. D.; Bures, M. G.; Martin, Y. C. 3_D property-based precursor selection for combinatorial library construction. National Meeting of the American Chemical Society, Chicago, IL; American Chemical Society: Washington, DC, August 20−25, 1995; Abstract 210.

CI950169+