Structural Diversity of Small Molecule Libraries

Anne Marie Munk Jørgensen[†] and Jan T. Pedersen*,[†]

Acadia Pharmaceutials A/S, Fabriksparken 58, 2600 Glostrup, Denmark

Received August 10, 2000

A novel method for assessing structural diversity is presented. Maximum common subgraph identity is used as the measure of similarity between two chemical structures. A conditional probability treatment of simmilarity distributions for libraries of chemical structures is used to define diversity. This evaluation method together with the evaluation of traditional physicochemical properties is used to assess a large number of chemical libraries and to understand structural differences between these.

INTRODUCTION

" *How do we optimize a given library of small molecules to provide an optimal set of both novel and known pharmacophore hits in a High Throughput Screening (HTS) setting?*".

Taking a purely 2D-structural approach, we have compared chemical graphs and compiled similarity/dissimilarity distributions for libraries of chemical structures. A conditional probability formalism is used to evaluate existing libraries and to build a screening library of currently 160 000 compounds which seeks to increase the hit rate toward known and novel GPCR receptor targets.

We are using a fast graph algorithm to compare 2D chemical graphs and identify the maximum common subgraph using the Bron and Kerbosh[1] clique finding algorithm. We also show that the same algorithmic machinery can be used for the evaluation of other structural parameters and features such as chirality and the identification of internal degrees of freedom.

The conditional probability formalism is ideal for the evaluation of chemical libraries when building both general purpose and focused screening libraries.

The evaluation of a library has been fully automated and results in a report describing the profile of the library. In this report we include the simmilarity/dissimilarity distributions and structural characteristics together with information about physicochemical parameters such as log P, pK_a, the number of hydrogen bond donors/acceptors, and the content of functional groups.

1. Structure Representation. It is trivial to make visual comparisons of small molecules using human intuition and pattern recognition. While the human eye can easily handle a limited number of 3D coordinates, this representation is difficult to handle in a computer, and many attempts have been made to reduce the complexity of molecular structure and the information contents of the Cartesian coordinates.

An excellent review of structure representation methods can be found in the seminal work by Peter Willett.[2] The structure representations can be roughly classified in four types of representation: *systematic nomenclature* (IUPAC), *fragmentation codes* (bit strings), *line notation* (SMILES, WLN), and *connection tables*.

The most popular description used which allows fast evaluation of any type of property is the classical bit-string encoding of information.[3] The bit-string encoding can be implemented to different levels of sophistication. Bit-string encoding is typically called "fingerprinting" because it generates a binary fingerprint of the object that it is supposed to describe. Bit-string encoding of properties allows for fast comparison and evaluation in a computer but possesses inherent disadvantages such as *information loss*, *no information weighting*, *irreversible encoding*, *coupled information*, and *inflexible encoding*.

Graph-theoretical algorithms have traditionally been used to identify common substructures or *graph isomorphisms*,[2,4−7] but these methods, which make a direct comparison between chemical structures, have not previously been used to statistically compare sets of molecules.

2. Similarity. Many similarity measures have been developed, and a review of such measures can be found in refs 2, 4−6, 8, and 9.

Considering the abstract concept of similarity measures, one can sort them in four different classes.[8]

Class 1: Distance Measures. These measures express the similarity or difference between two molecules as the Euclidian distance between two molecules in a multidimensional descriptor space, where the descriptors are variables that describe the shape, composition, and chemical properties of the molecule.

Class 2: Association Measures. Association measures are used to describe the similarity between two binary descriptor strings. The widely used Tanimoto[8] index is an example of an association measure.

Class 3: Correlation Measures. These measures are similar to the distance measures but describe the statistical significance of corellations between two sets of variables instead of measuring the actual distance between the sets of variables.

Class 4: Probabalistic Measures. These types of measures take into account the frequency of occurrence of observed variables in datasets.[10]

* Corresponding author. Phone: +45 36 30 13 11. FAX: +45 36 30 13 85. E-mail: jatp@lundbeck.com.

[†] Current address: H. Lundbeck A/S, Ottilia vej 9, DK-2700 Valby, Denmark.

10.1021/ci000111h CCC: $20.00 © 2001 American Chemical Society
Published on Web 02/15/2001

Lately, neural network methods have been developed that allow an encoding of weighted information.[11] A structural data input can be related to physical−chemical microscopic/macroscopic parameters (effects) or to abstract effects such as "drug-likeness".[11] Although, the main disadvantage by this type of method is that it is difficult to extract the structural information from a neural network analysis, the methods show promise. This type of similarity measure can be considered to belong to class 4 above.

In this paper we present another type of class 4 similarity measure. We use a statistical treatment of graph similarity as a measure for similarity and diversity within chemical libraries.

METHODS

1. Definition of Structural Similarity and Diversity. Structural Similarity. The similarity between two structures is calculated as the average identity to each of the two parent structures:

$$S_{ij} = \frac{1}{2}\left(\frac{N_{\mathrm{as}}}{N_i} + \frac{N_{\mathrm{as}}}{N_j}\right) \tag{1}$$

where N_{as} is the number of connectivities (bonds) in the maximal, connected, subgraph identified by comparing structures i and j. N_i and N_j are the numbers of bonds in structures i and j.

This measure of similarity has some obvious advantages. A similarity of 100% is only obtained when the two structures compared are exactly identical. Second, the same similarity is obtained when molecule i is compared to molecule j and molecule j is compared to molecule i.

Comparing a large molecule to a small one, where all of the small molecule is represented by a substructure of the large molecule, results in intuitively correct similarities (S_{ij}). By which we mean comparing ethylene to naphthalene ($S_{ij} = (1/2)(1/11 + 11/11) = 12/22 = 0.55$) is less significant than comparing benzene to naphthalene ($S_{\mathrm{ij}} = (1/2)(6/11 + 6/6) = 17/22$.

Library Diversity. For an ensemble of structures the diversity (D) is defined as the probability distribution of dissimilarities:

$$P(D) = \frac{\sum_{p=1}^{n} D_p}{N} \tag{2}$$

where D_p is a structural 2D dissimilarity of a particular value and is defined as $D_{ij} = 1 - S_{ij}$. D_{ij} is the dissimilarity between two molecules i and j. N is the total number of structure comparisons performed. In practice we use $P(S)$, which is the mirror image distribution of $P(D)$.

Library Evaluation. The question we are interested in evaluating in a library analysis is, "What is the probability that a given library will contain a compound that is active against a specific receptor". To answer this question, we would need to generate a biological profile for a large number of random compounds. The probability we are evaluating is

$$P(R_i|L) = \frac{N_{Ri}}{N_{\mathrm{total}}} \tag{3}$$

where $P(R_{\mathrm{i}}|L)$ is the probability of finding a compound that shows a biological activity against a specific receptor R_i in a given library L. $N_{\mathrm{R_i}}$ is the number of compounds that show the desired biological activity out of all the compounds in the library (N_{total}).

For a virtual library (and for many real libraries) it is not possible to perform the above biological evaluation, and in that case we would like to get an estimate of $N_{\mathrm{R_i}}$ for a large number of receptors.

Assuming that a collection of all common and experimental drugs embodies all the currently known pharmacological activities, it is possible to write

$$R = \sum_{i=1}^{N} R_i \tag{4}$$

where R represents all known receptor activities and N is the number of known common drugs with unique activities. We here assume that we have filtered our library such that close analogues have been eliminated in order to eliminate similar biological activities. On the other hand it is important to state that two structurally distinct HIV protease inhibitors will be retained in R since their molecular binding mode is likely to be different.

The chance of finding a biological activity in a given library can then simply be evaluated as

$$P(R|\ L) = \frac{N \cap \mathbf{N_{total}}}{\mathbf{N_{total}}} \tag{5}$$

or it can simply be formulated as the fraction of compounds in a given library that exists in our database of biologically active compounds relative to the total size of the library. The main problem with this formulation is that N is a relatively small number (3000−5000) and covers an order of magnitude fewer pharmacophore entities. The prediction can be improved by implying that structural similarity infers biological activity. That is,if two compounds are structurally similar, they are likely to have similar biological activities. This means that we can rewrite eq 5 in terms of similarities.

$$P(R_i|\ L) \approx \frac{N_{\mathrm{Sc}}}{N_{\mathrm{cmp}}} = \frac{N_{\mathrm{Sc}}}{{N_{\mathrm{total}}}^2} \tag{6}$$

where N_{Sc} is number of structural comparisons where the similarity is higher than a given cutoff c. N_{cmp} is the total number of comparisons. In reality one only has to make $(1/2){N_{\mathrm{total}}}^2$ comparisons since $S_{ij} = S_{ji}$ according to eq 1.

From eq 6 it can be seen that if the assumption that structural similarity does not imply similar biological activity (i.e. $N_{\mathrm{S_c^{real}}} \ll N_{\mathrm{Sc}}$), the probability $P(R_i|L)$ will be overestimated. However, we do not consider this assumption to be controversial since it also forms the basis of all traditional and well-proven structure−activity relationship (SAR) and quantitative SAR (QSAR) methods.

The above formula does not provide any indication of the probability of finding a structurally novel hit in a given library or a hit against a new receptor target. If one wishes

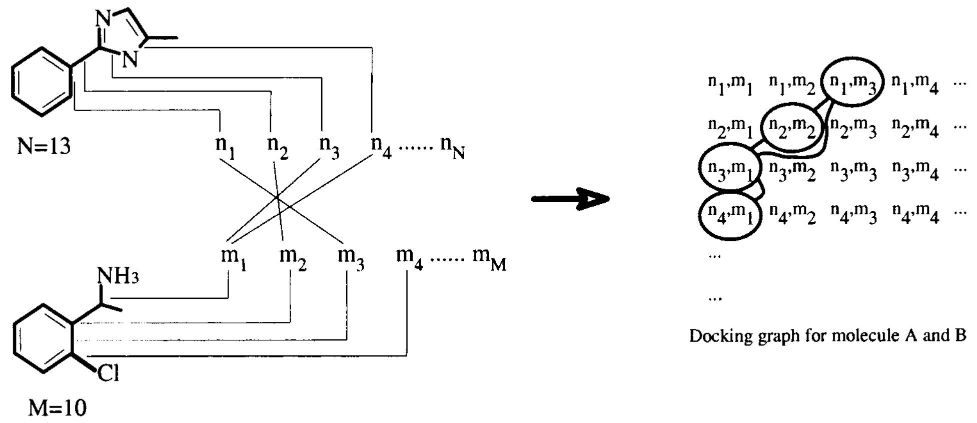

N
N
N=13
n_1 n_2 n_3 n_4 ...... n_N
m_1 m_2 m_3 m_4 ...... m_M
NH3
Cl
M=10
n_1,m_1 n_1,m_2 n_1,m_3 n_1,m_4 ...
n_2,m_1 n_2,m_2 n_2,m_3 n_2,m_4 ...
n_3,m_1 n_3,m_2 n_3,m_3 n_3,m_4 ...
n_4,m_1 n_4,m_2 n_4,m_3 n_4,m_4 ...
...
...
Docking graph for molecule A and B

Figure 1. The basis of the generation of the comparison or *docking* graph. The docking graph consists of all pairs of identical bonds in two structures *A* and *B*. A node n_{ij} is defined by two bonds where the connecting atoms are identical and the bond order is the same. Two nodes (n_i,m_i) and (n_j,m_j) in the docking graph will be connected by an edge if the *shortest* distance between n_i and n_j in the first structure is equal to the *shortest* distance between m_i and m_j in the second structure. The *distance* between two bonds in a 2D graph is defined as the number of bonds one has to pass through to get from bond n_i to n_j. A clique in the docking graph is a subgraph where all nodes are connected to each other by edges. The largest common substructure is defined by the largest clique. In the example above nodes (n_1,m_3),-(n_2,m_2),(n_3,m_1) constitute a clique since all nodes are connected to each other. Only part of the complete docking graph is shown for molecules A and B.

to evaluate the probability that a new library will provide a structurally novel hit, one needs to evaluate the *diversity* or breadth of a given library. Chemical space is infinite, and the only limitations to the breadth of the library is the desired pharmacological, toxicological, and pharmacokinetic profile of a given drug together with the desired molecular weight range. This same “gedanken” experiment has led to the development of Lipinski’s acclaimed Rule of Five.[12]

To monitor the width (diversity in our strict definition) of an HTS library, it is necessary to evaluate the probability:

$$P(i_L|L_r) \approx \mathrm{N}_{\mathrm{Sc}}/N_{Lr\mathrm{total}} \qquad (7)$$

$P(i_L|L_r)$ is the probability that compound *i* within the library *L* will exist in some large reference library L_r consisting of random structures. N_{Sc} is the number of structural comparisons where the similarity is higher than a given cutoff *c* and $L_{r\mathrm{total}}$ is a suitable random reference distribution.

The relevant reference distribution is a library of randomly generated chemical structures which satisfy the size, pharmacological, toxicological, and pharmacokinetic profile of a common drug.

Any exploratory library must therefore strive to find as wide a distribution for *L* as possible within the physiologically relevant boundaries (*L* approaching L_r).

2. Graph Searching Algorithm. Definition of Graphs. Two structure graphs are generated for the two molecules that are to be compared (Figure 1). It follows from this figure that each bond in the chemical structure of a molecule becomes a *node* in the graph. The bond is characterized by the atom types of the end atoms and the bond order. All nodes are connected by *edges* describing the smallest through-bond distance, between each pair of bonds in the molecule. The distance is 0 if the bonds are connected, it is 1 if there is a single bond between, etc. In the program, this shortest distance between two bonds is determined by use of an *all pairs shortest path* algorithm.

A comparison graph is defined on the basis of two structural graphs. All combinations of identical nodes within the two graphs become nodes in the comparison graph. If node *a* in graph *A* is identical to node *b* in *B*, the pair (*a*,*b*) becomes a node in the comparison graph. Further, two nodes, (*a*,*b*) and (*c*,*d*) in the comparison graph will be connected by an edge if the through-bond distance connecting *a* and *c* in graph *A* is equal to the through-bond distance connecting nodes *b* and *d* in graph *B*.

When the entire comparison graph has been constructed, the largest common subgraph is identified using a clique finding algorithm developed by Bron and Kerbosh.[1] This algorithm uses a combination of recursive backtracking and a “branch and bound” method to eliminate searches that will not lead to cliques. The recursive procedure used here is self-referential: finding a clique of length *n* is accomplished by first finding a clique of length $n - 1$ and finding another node that is connected to all the nodes in that clique. The branch and bound technique makes use of rules that allow us to determine in advance certain cases for which possible combinations of nodes and edges that will never lead to a clique.

The size of the maximal clique corresponds to the number of bonds in the maximum common substructure for the two molecules. By using eq 1, defined above, this value is used to calculate the similarity between the two molecules. It is important to note that the maximum common substructure may not be connected (Figure 2).

The substructure search algorithm, presented here, is analogous to the structural comparison method, published by Takahashi et al.[6] However, important differences exist between the two methods. The present method makes the structural comparison on an all atom level, resulting in an absolute comparison. In the presented method nodes represent bonds characterized by bond order and atom types. Intuitively, a more natural representation of the chemical structure would have atoms as nodes. This results in a less restricted representation of the chemical structure, approximately doubling the number of possible combinations of similar nodes, dramatically increasing the size of the comparison graph. The restricted representation algorithm results in fewer nodes in the comparison graph, making the clique finding a less demanding computational task. Additionally, in the presented method we only consider the

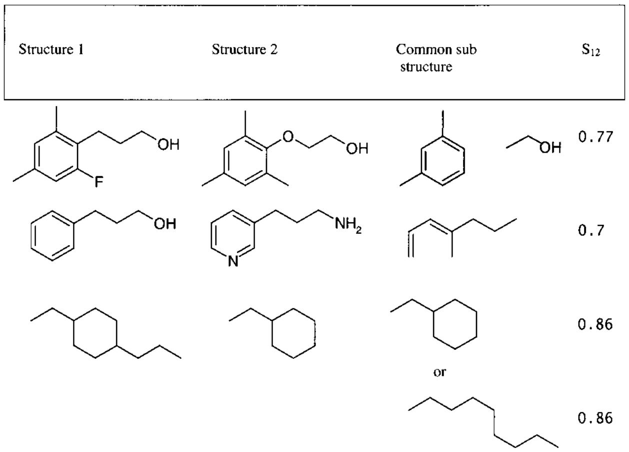


Figure 2. Three examples of structure comparisons by use of the common substructure search algorithm. The similarity between the structures 1 and 2, S_{12}, is defined as $S_{12} = (n/n_1 + n/n_2)(1/2)$, where n is the number of bonds in the common substructure and n_1 and n_2 are the numbers of bonds in structures 1 and 2, respectively. The common substructure will in some cases consist of more than one structural fragment, as demonstrated in A. It is possible that more than one maximal clique exists with the same size, resulting in the same value of S_{12}. An example of this is shown in panel C. In these cases the program selects the clique with the fewest number of atoms included, which corresponds to a selection of the most connected common substructure. In example C, the program would, therefore, choose the first solution.

shortest distance between two bonds instead of the full set of all distances between the bonds, also reducing the number of comparisons and decreasing the total computational effort.

Before comparison, structures are preprocessed to allow for multiple conjugation of aromatic ring systems.

3. Evaluation of Structural Properties. The chemical graph and the similarity search algorithm, described above, can also be used to evaluate structural properties such as stereochemistry, internal degrees of freedom (number of rotatable bonds), and distribution of functional groups within a given chemical structure.

Stereogenic Centers. To evaluate the existence of stereogenic centers, it is necessary to recursively evaluate each branch from all sp^3 hybridized carbon atoms in a given molecule. For each of these atoms four substructure graphs are constructed, as shown in Figure 3, each containing a substituent atom and all the atoms and bonds connected to the substituent. Substituent atoms in rings are treated as a special case; If two or more substituents are a member of the same ring, the substituent graph is split at the connection point to the atom that is currently investigated (Figure 3).When four substructure graphs have been successfully constructed, all possible pairs are compared by using the similarity search algorithm, described above. If no pairs of substructures are identical, a stereogenic center has been identified.

Rotatable Bonds. All single bonds that satisfy the following criteria are identified as rotatable bonds:

(a) The heavy atoms (non-hydrogen atoms) A−B connected by a single bond, must both be connected to a second atom (C,D) C−A−B−D. This second atom may be a hydrogen atom.

(b) C−A or B−D must not be a triple bond unless the triple bonded atom is connected to another atom.

(c) The bond A−B must not be part of a ring.

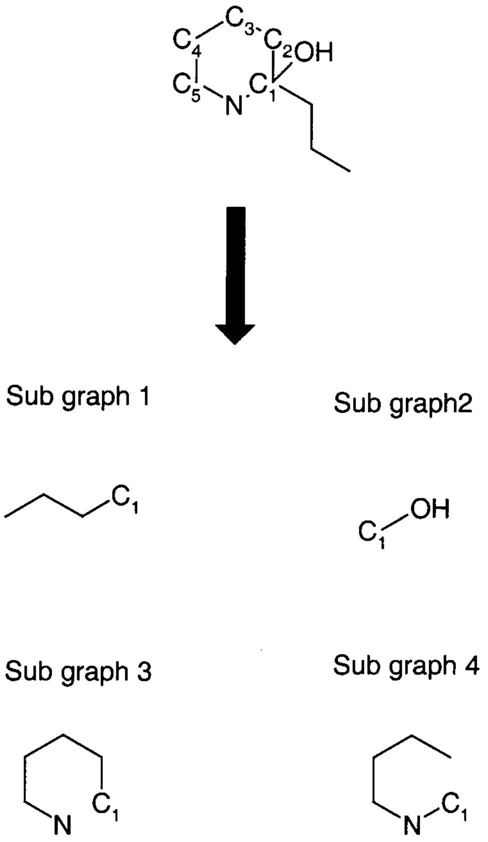


Figure 3. Demonstration of the generation of the substructure graphs used in the algorithm, which examine for chirality. The example shows the examination of atom C_1. When four substructure graphs have been constructed, all possible pairs are compared by using the similarity search algorithm, described above. If no identical pair is found, C_1 is identified as a stereogenic center.

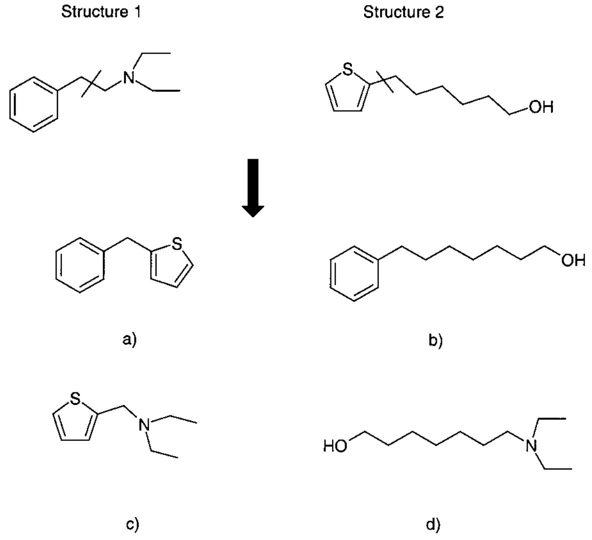
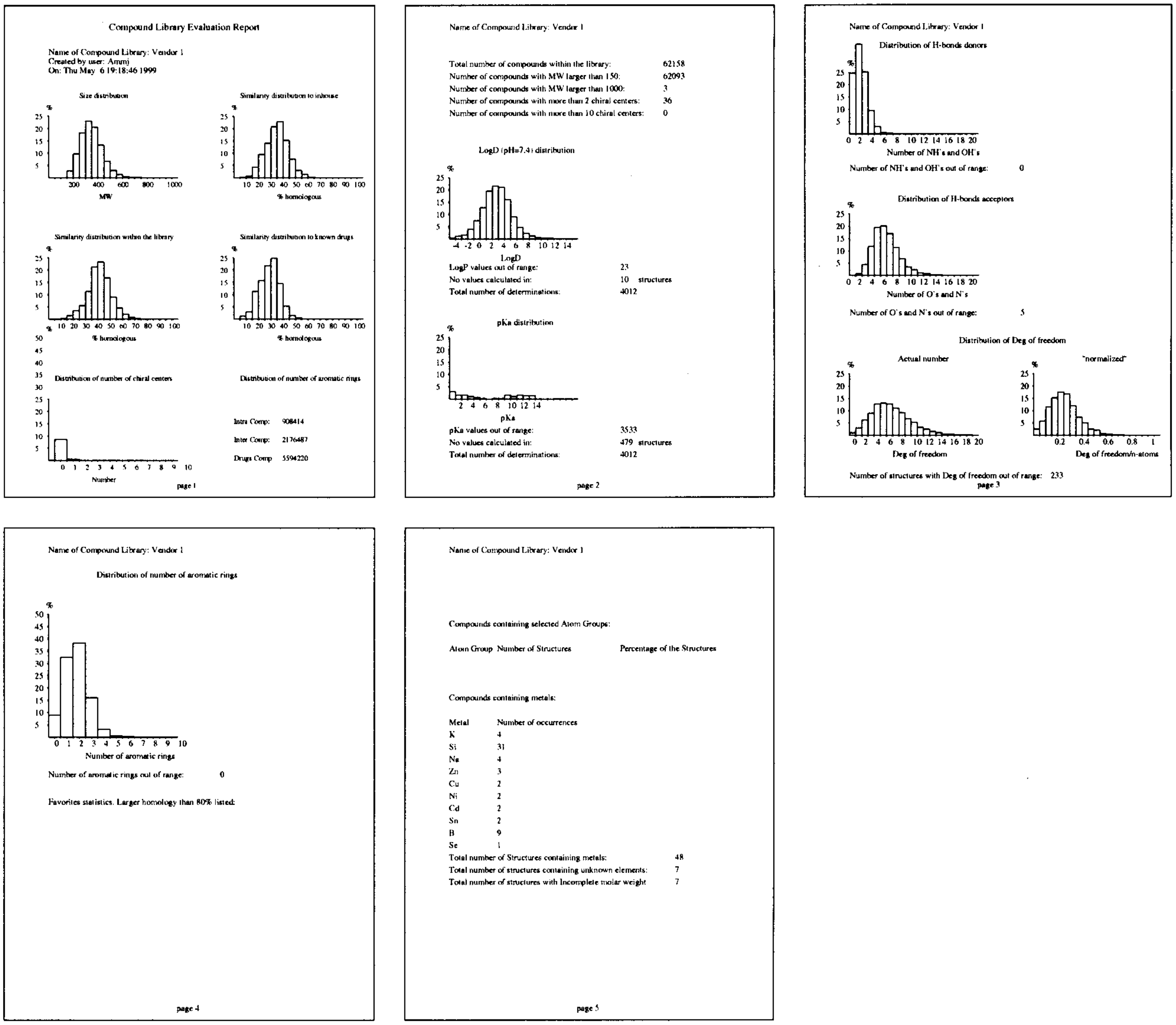


Figure 4. Random library generation by recombination of structures 1 and 2. For each pair of structures, four new structures are made by the shown recombination process. Thus, each of the molecules are cut into two, by randomly selecting and breaking a single isolated bond. This is indicated by the bar. The two halves are then recombined with each of the halves from the other molecule. To reduce the number of compounds, only a randomly chosen subset of these combinations was carried out. Furthermore, a selection procedure was included to ensure that the size distribution in the original library was retained within this "random" library.

These properties are checked for each bond in a given molecule.



Compound Library Evaluation Report

Name of Compound Library: Vendor 1
Created by user: Ammj
On: Thu May 6 19:18:46 1999

Intra Comp:	908414
Inter Comp:	2176487
Drugs Comp	5594220

page 1

Name of Compound Library: Vendor 1

Total number of compounds within the library:	62158
Number of compounds with MW larger than 150:	62093
Number of compounds with MW larger than 1000:	3
Number of compounds with more than 2 chiral centers:	36
Number of compounds with more than 10 chiral centers:	0

LogP values out of range:	23
No values calculated in:	10 structures
Total number of determinations:	4012

pKa values out of range:	3533
No values calculated in:	479 structures
Total number of determinations:	4012

page 2

Name of Compound Library: Vendor 1

Number of NH's and OH's out of range: 0

Number of O's and N's out of range: 5

Number of structures with Deg of freedom out of range: 233

page 3

Name of Compound Library: Vendor 1

Number of aromatic rings out of range: 0

Favorites statistics. Larger homology than 80% listed:

page 4

Name of Compound Library: Vendor 1

Compounds containing selected Atom Groups:

Atom Group Number of Structures Percentage of the Structures

Compounds containing metals:

Metal	Number of occurrences
K	4
Si	31
Na	4
Zn	3
Cu	2
Ni	2
Cd	2
Sn	2
B	9
Se	1

Total number of Structures containing metals:	48
Total number of structures containing unknown elements:	7
Total number of structures with Incomplete molar weight	7

page 5

Figure 5. Library evaluation report for the virtual library from vendor 1. This is a general example of a library evaluation report. Page 1 lists general 2D structure comparison distributions. Size distribution and a distribution of the number of chiral centers are shown. Page 2 lists general properties of the library, such as size and molecular weights over 150 and 1000. Distribution of physical chemical properties, such as log D and pK_a, is also listed. Page 3 lists distributions of steric properties such as H-bond donors and H-bond acceptors and the distribution of the number of internal degrees of freedom. Page 4 contains a distribution of the number of aromatic rings and a list of hits from a comparison to an in-house list of favorite structures. The latter list is used to identify new leads in existing in-house projects. Page 5 lists rare elements identified in the library.

4. Evaluation of Physical−Chemical Properties. Distributions of the log P and pK_a values are also included in the library analysis. These values are all calculated using the empirical program xpred.[13] PALLAS calculations are based on a PLS analysis of fragment contributions.[13]

Functional Groups. Functional groups are identified by using the above-described structure searching algorithm to identify aromatic rings, H-bond donors/acceptors, and "favorite" substructures defined in a substructure library. Favorite substructures are pharmacophore scaffolds which are used in in-house chemistry projects and which have repeatedly been identified in screening programs.

5. Structural Libraries. To determine N_{Sc} in eq 2, we make use of an approach similar to that described by Lipinski et al.[12] All structures in the MDDR database that contain an INN or USAN record are extracted.

In-House Library. The in-house library is a set of approximately 160 000 compounds collected from a diverse set of sources. The compounds come from more than 10 commercial suppliers and more than 50 academic laboratories. To optimize diversity of the in-house library, compounds are included in this library if less than 20% of the compounds in a new library have more than 50% structural homology. These cutoffs have been chosen arbitrarily from experience.

Reference Library. The random reference library is generated by randomly combining fragments of structures from the common drug library. For each combination of structures within the library, four new structures are generated by a genetic algorithm in which each of the structures are cut into two by randomly selecting and breaking an isolated single bond. The two halves are then recombined with each of the halves from the other structure. To reduce the number of generated structures, only a randomly sampled subset of these combinations was performed. Furthermore, a selection procedure was included to ensure that the size distribution in the original library was retained within the reference

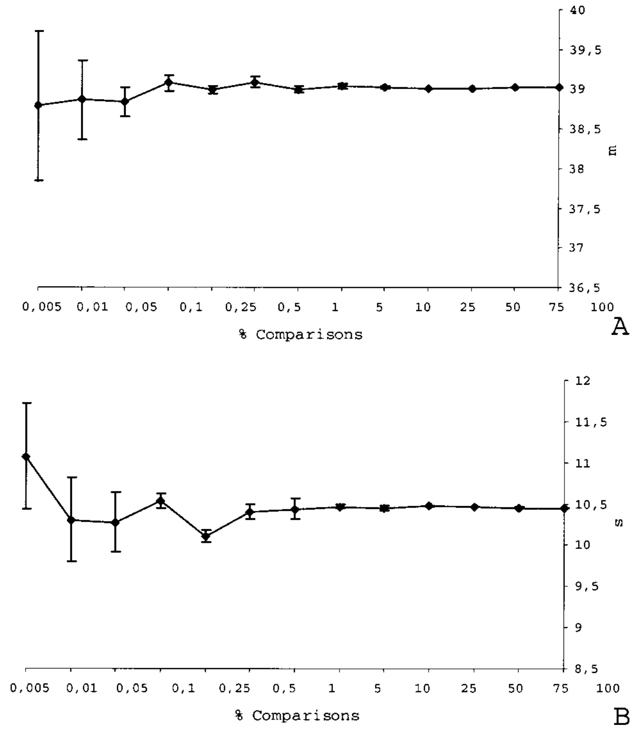


Figure 6. Relationship between the similarity distribution and the percentage of the total number of comparisons made, all comparisons selected randomly. Calculations were performed on two sets of 1000 compounds selected by random from our in-house library of 160 000 compounds. For each of the similarity calculations a normal distribution was fitted to the similarity distribution and was characterized by the mean (*m*) and the spread (*s*) of the fitted distribution. For each percentage a number of similarity calculations were carried out to see how these varied. For 5% and all percentages below 5%, 10 calculations were carried out for each percentage; for percentages above 5%, 5 calculations were performed, except for 100%, where only one calculation was made. The variation of the mean *s* and *m* values, as a function of the percentage of comparisons made, are shown in A and B, respectively.The standard deviations in *s* and *m* for comparisons carried out with the same number of comparisons are indicated by the error bars.

library. Otherwise, the generated library would contain a large amount of very small molecules (molecular weight less than 50) and a corresponding large amount of high-molecular weight molecules. The described procedure resulted in a reference library consisting of 7025 structures. The recombination process is illustrated in Figure 4.

RESULTS

1. Library Evaluation. To be able to assess the structural quality of a given library, we compute a number of distributions and characteristic parameters for these distributions. These distributions and characteristic numbers are formed into an *evaluation report*. An example of this report is shown in Figure 5.

Three probability distributions are computed for the evaluation of structural properties of the library under investigation: similarity of the library to itself, providing a view of the internal diversity of a given library together with a comparison to the random reference library providing an estimate of $P(i_L|L_r)$; similarity of library to library of common drugs, providing an idea of pharmacologically proven relevance of scaffolds in the new library (this is an estimate of $P(R_i|L)$); similarity of library to in-house library (from this distribution the contribution to the in-house library diversity can be calculated).

Table 1. Evaluation of 20 Libraries[a]

library	no. of compds	m	s
1	14 789	38.6	10.9
2	17 055	36.9	12.0
3	20 104	38.1	12.3
4	8 800	40.0	13.6
5	8 800	38.5	11.6
6	8 800	38.9	11.2
7	3 015	36.0	11.7
8	4 621	39.2	13.2
9	4 737	39.4	12.8
10	62 158	37.9	11.7
11	21 398	37.9	12.5
12	16 500	38.6	11.6
13	2 231	36.9	11.4
14	1 525	36.4	11.2
15	4 102	40.4	19.4
16	20 000	44.5	13.1
17	4 000	45.0	13.1
18	601	48.9	16.0
"common" drugs	4 478	33.8	11.9
random	7 025	30.5	9.5

[a] Each library was compared to itself by use of the similarity search algorithm described in the text. A normal distribution was fitted to each resulting similarity distribution, and the spread (*s*) and mean (*m*) values are listed in the table. The "random" library is made by recombination of structural fragments within library 15 in the table. The recombination algorithm is described in the text.

2. Extent of Comparisons. The computational time for comparing all structures within a library is $O(n^2)$, where n is the number of structures in the library. When the size of the library under examination increases, comparing all structures becomes a computationally demanding task. Therfore, we have investigated how the similarity distribution varies as a function of the number of comparisons made. Similarity calculations with 0.005, 0.01, 0.05, 0.1, 0.25, 0.5, 1, 5, 10, 25, 50, 75, and 100% of the total number of comparisons were performed. Comparisons were randomly selected from the set of all (n^2) comparisons. For each percentage a number of similarity calculations were performed to determine the accuracy of the experiment. For all percentages below 5%, 10 calculations were performed; for percentages above 5%, five calculations were performed, except for 100%, where only one calculation was made. The two test libraries contained 1000 structures, corresponding to 10^6 comparisons for the full set. For each of the similarity calculations a normal distribution was fitted to the similarity distribution and was characterized by the mean (*m*) and the spread (*s*) of the fitted normal distribution. The variations of the mean (*s*) and (*m*) values as a function of the percentage of comparisons made are shown in Figure 6A,B. The standard deviations in *s* and *m* are indicated by the error bars. Only 1−5% of the total number of comparisons need to be made in order to obtain a well-defined distribution as indicated by the small error.

This analysis was carried out for libraries of varying size to determine how variation in the standard deviations of *m* and *s* changed when the number of structures within the libraries changes.

To be able to use *m* and *s* as characteristic parameters describing a given library, these parameters need to be determined with the same accuracy. As seen in Table 1 the variations in *s* and *m* values are small, and for this reason the number of comparisons has to be adjusted to the size of the library. We performed the calculation described above for four libraries containing 100, 500, 1000, and 10 000 structures. A threshold value of 0.03 and 0.06 was chosen for the standard deviation of the *s* and *m* values, respectively. To get standard deviations below these values, we had to include approximately 100, 25, 10, and 0.1% of the total number of comparisons for each of the libraries.

3. Evaluation of Libraries. A large number of compound libraries from different vendors was analyzed by use of the described similarity searching algorithm. Table 1 shows the results of 20 calculations where different libraries were compared to themselves; A normal distribution was fitted to the resulting similarity distribution, and the spread (*s*) and mean (*m*) values are listed in Table 1. As seen in the table, most "historical" collections, those are libraries 1−16 in the table, have mean values in the range of 36−39 and a spread in the range of 10−13. In contrast, compound libraries resulting from combinatorial chemistry, these being libraries 17−18, have mean values above 45. The spread is almost the same for these libraries compared to the collected libraries. The "random" library is the reference library described previously. The random library has a mean value of 30.47, indicating that it may be difficult to find libraries with mean values much smaller than 30. Perhaps surprising, the compounds within the common drug database have a low mean value of 33.75, indicating that the common drugs dataset corresponds to a structurally diverse library of compounds.

4. Evaluation of Functional Group Distribution. Four different kinds of compound libraries were examined for contents of functional groups. These libraries are as follows: (a) the common drug library; (b) a library resulting from combinatorial chemistry; (c) the "random" library, described above; (d) a typical collected ("historical" collection) library. The results of this analysis are shown in Figure 7. As seen in this figure, a number of functional groups are well-represented in the common drug library; this is especially true for carboxylic acid (1), keto (22), formamido (23), amine (29), thioether (30), ether (31) and ester (32), where the numbers in parentheses refer to Figure 7. When the common drug library is compared to a library resulting from combinatorial chemistry, it appears that the formamido group (24) is overrepresented, as seen in panel B. Both the random library and the collected library have functional group distributions which are much closer to the distribution in the common drug library, as seen in panels C and D, respectively. The "random" library is constructed from building blocks within the common drug library, and therefore, the functional group distribution will to a certain extent reflect the functional group distribution within this library.

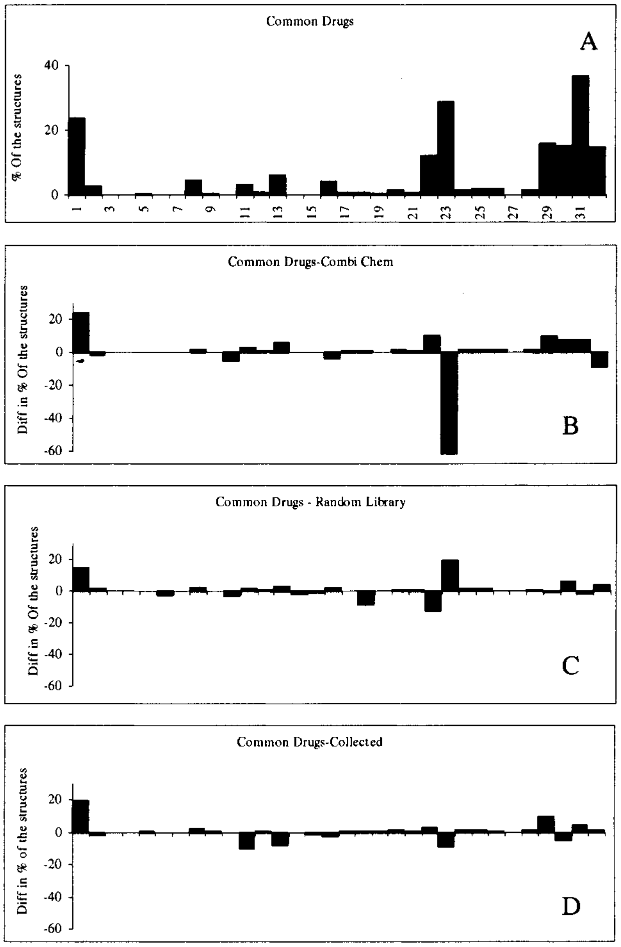


Figure 7. Distribution of functional groups in various libraries: (A) the common drug library; (B) a library resulted from combinatorial chemistry; (C) the "random" library; (D) a collected library. In A, the histogram shows the percentage of the structures having a certain functional group. In B−D, the histograms show the difference between the percentage of the structures in the common drugs library and the percentage of the structures in the actual library, having a certain functional group. The functional groups are as follows: (1) carboxylic acid; (2) nitril; (3) diazonium salt; (4) azide; (5) isocyanide; (6) isocyanate; (7) isothiocyanate; (8) dimethylamine; (9) trimethylamine; (10) nitroso; (11) nitro; (12) nitrate; (13) hydroxylamine; (14) anhydride; (15) trichlor; (16) triflur; (17) phosphate; (18) aldehyde; (19) phosphinate; (20) phosphite; (21) sulfate; (22) keton; (23) formamido; (24) sulfino; (25) sulfite; (26) sulfone; (27) thionyl chloride; (28) sulfoxide; (29) amine; (30) thioether; (31) ether; (32) ester.

DISCUSSION AND CONCLUSIONS

The graph defined structural similarity measure (*S*) is a direct measure of structural diversity and complements the use of keys or fingerprint encoding of chemical structure. A diversity measure (*D*) can be derived directly from the similarity definition that has a direct structural meaning. In this study we have focused on exploring a new formalism that lends itself to objective evaluation of structural diversity. Although, no comparison to other methods is presented, it appears to perform similarly to fingerprint clustering methods in a number of test cases, in that it generates similar clusters and clusters similar structures together.

The conditional probability formalism is a good way of evaluating structural libraries and of formulating specific questions and requirements to a given library. That is, if one seeks to develop a screening library which has properties similar to those found in an already known selection of structures, a screening library can be biased toward the desired pharmacophores without excluding new pharmacophore scaffolds.

Since the method has a statistical basis, it is unnecessary to make all n^2 structural comparisons; on average one only needs to make 0.5−1.0% of the total comparisons. This makes the method attractive because a large library ($10^{5-}10^{6}$ compounds) can be evaluated quickly.

The conditional probability formalism can be used to evaluate diversity within a given library and between libraries. That is, one can directly compare the quality of a number of libraries with a common reference point. In this case the common reference point is a random reference library.

When using the conditional probability formalism to evaluate a library of experimental and approved drugs, one finds that such a library is highly diverse. The only library we have encountered which is structurally more diverse than the drug library is our randomly assembled reference library. This suggests that there is no structural feature (pharmacophore) which is more "druglike" than any other. Features that discriminate "druglike" from "nondruglike" compounds are purely physicochemical. The physicochemical properties are a feature of the detailed chemical structure; that is, the removal or addition of a single group or atom from a chemical structure can dramatically alter the macroscopic properties of a given drug molecule.

It is difficult to make a chemical library which is structurally more diverse than the common drugs library, suggesting that a collection of experimental and approved drugs would make a diverse screening library. This library represents many years of chemical intuition compressed into one library where each compound represents the end-point of a large medicinal chemistry effort.

REFERENCES AND NOTES

(1) Bron, C.; Kerbosh, J. Finding all cliquies of an undirected graph. *Commun. ACM* **1973**, *16*, 575−577.

(2) Willett, P. *Similarity and clustering in chemical information systems*; Research Studies Press, John Wiley & Sons, Inc.: New York, 1987.

(3) Craig, P. N.; Ebert, H. M. Eleven years of structure seraching using the skf (smith, kline and french) fragmentaton codes. *J. Chem. Doc.* **1969**, *9*, 141−146.

(4) Johnson, M. A., Magiora, M., Eds. *Concepts and application of molecular similarity*: John Wiley & Sons, Inc.: New York, 1990.

(5) Cone, M. M.; Venkataraghavan, R.; McLafferty, F. W. Molecular structure comparison program for the identification of maximum common substructures. *J. Am. Chem. Soc.* **1977**, *99*, 7668−7671.

(6) Takahashi, Y.; Sukekawa, M.; Sasaki, S. Automatic identification of molecular similarity using reduced-graph representation of chemical structures. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 639−643.

(7) Read, R. C.; Corneil, D. G. The graph isomorphism diease. *J. Graph. Theory* **1977**, *1*, 339−363.

(8) Willett, P.; Winterman, V. A comparison of some measures for the determination of intermolecular structural similarity. *Quant. Struct.-Act. Relat*. **1986**, *5*, 18−25.

(9) Varkony, T. H.; Shiloach, Y.; Smith, D. H. Computer-assisted examination of chemical compounds for structural similarities. *J. Chem. Inf. Comput. Sci*. **1979**, *2*, 104−111.

(10) Adamson, G. W.; Bush, J. A. A comparison of the performance of some similarity and disimilarity measures in the automatic classification of chemical structures. *J. Chem. Inf. Comput. Sci*. **1975**, *15*, 55−58.

(11) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and non-drugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(12) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Freeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Del. Rev.* **1997**, *23*, 3−25.

(13) *CompuGrug, Xpred, PrologP, pKalc: Programs for the calculation of LogP and pka values*, versions 5.11 and 3.21; 1992, 1996.

CI000111H