*First International Electronic Conference on Synthetic Organic Chemistry (ECSOC-1), www.mdpi.org/ecsoc/, September 1-30, 1997*

[F0002]

# Design and Analysis of Diverse Screening Libraries

by Dirck Lassen

TRIPOS GmbH, Martin-Kollar-Str. 13, D-81829 München, Germany. E-mail: dlassen@tripos.com

*Received: 15 July 1997 / Uploaded: 29 August 1997*

**Abstract**

Screening very similar compounds is wasteful. The paradigm that structurally similar molecules exhibit similar physicochemical and biological properties (1) can be used to reduce existing screening pools to small representative subsets which enhances the efficiency to find novel lead structures.

The stategy is applied to two examples: a combinatorial library based on a single scaffold and an existing structurally diverse compound database. While the combinatorial library contains 97 % redundant structures, for the compound database 62 % of the structures are redundant and can be filtered out leading to optimal diverse libraries. The obtained diverse subsets enhance the structural diversity of the Maybridge database.

## 1. Introduction

Similarity considerations are nowadays widely used (2) to design a collection of optimally diverse compounds spanning a wide range of chemical and biological properties for synthesis and/or assay-screening. The *Similarity Principle*, explicitly stated by Johnson & Maggiora in 1990 (1), forms the conceptual basis for these investigations. *Structurally similar molecules should exhibit similar physicochemical and biological properties.* Two important conclusions can be derived:

1. Prediction of unknown biological or physicochemical properties is possible given a similarity relationship between two molecules.

2. A representative compound subset should cover the entire property space of a larger database. An optimal selection would include as few similar members as possible.

Thus, the use of very similar molecules for screening does not enhance the probability to find different types of biological activities, but by using structurally dissimilar molecules the probability for finding interesting leads is higher. Obviously a diverse compound library should sample a wide range of diverse, non-redundant compounds. However, not only the risk of missing a lead compound should be kept low, but also the total number of compounds.

The concept of redundancy can be illustrated using the simplified picture in Fig. 1. While a typical chemical database is characterized by well separated clusters of very similar compounds (e.g. variations of core structures), an optimal distribution of molecules will avoid this degree of redundancy and the loss of information. Such a database can be designed taking structural properties into account to determine the smallest acceptable distance between two molecules ("similarity radius"). Taking a large variety of compounds within a initial virtual library, an optimal procedure would select only dissimilar compounds outside this similarity radius (Fig. 1 bottom), leading to a more diverse subset, which increases the probability of finding lead structures. Moreover such a distribution may help to identify structurally different molecules being active on the same target, supposing that the interest lies in the differentiation of active vs. inactive compounds regardless of a quantification of activity. Subtle activity differences are subject of a subsequent lead refinement program based on similarity and analog libraries (3).

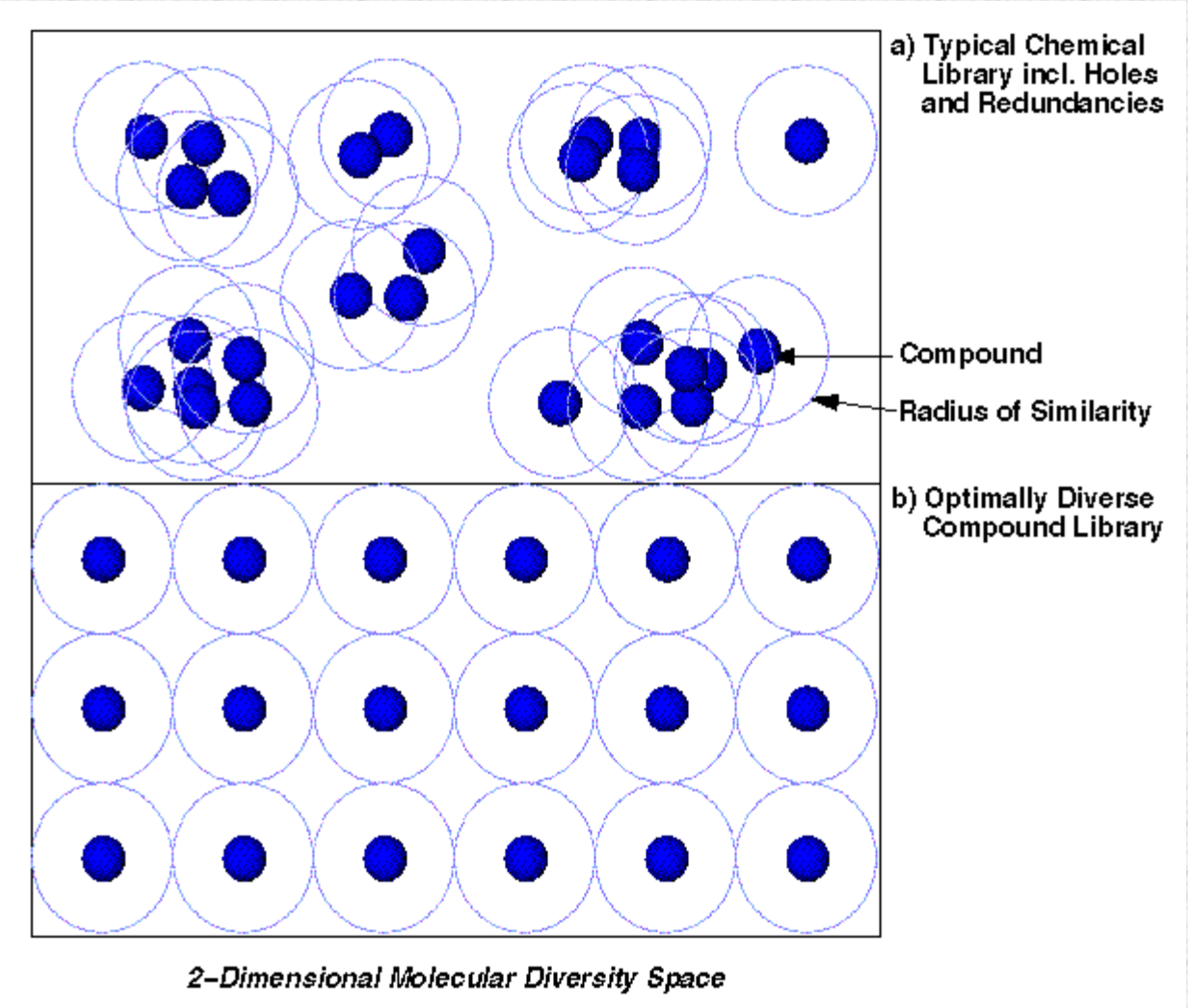



*Fig. 1: A simplified representation of a typical chemical compound library (upper part) and an optimal diverse library (lower part) shown as distribution of molecules (filled circles) in an arbitrary 2D molecular property space. The typical database consists of redundant structures (clusters of very similar molecules) and "holes", which are defined as uncovered regions in this property space. The dotted circles represent a radius of similarity, corresponding to the uniqueness or redundancy of molecules.*

Compound selections were carried out based on this dissimilarity concept using 2D fingerprint descriptors within a database containing structurally diverse templates and 55 biological target properties and within a virtual library of 10752 benzodiazepines based on Ellman's reaction scheme (4). In addition these studies demonstrate that the use of 2D fingerprints lead to a good representation of global molecular diversity.

## 2. Methods

All modeling and library design was done using the Tripos Molecular Diversity Manager, which is a set of integrated software tools that provides a means for faster and more efficient analysis of HTS data and structures. The Molecular Diversity Manager is composed of three key components. Legion (5) enables building and managing virtual combinatorial libraries of compounds. Unity (6) provides storing, searching and analysing chemical structures and Selector (5) does rational selection of compounds based on diversity and similarity. Automation of procedures was done using the SYBYL Programming Language (SPL). Data manipulations

were carried out using SYBYL's Molecular Spreadsheet (5) or Unity databases. 2D fingerprints were calculated as standard UNITY fingerprints.

## 2.1. Diversity metric

The design of diverse libraries relies on suitable metrics. Any metric used for measuring molecular diversity can be validated. The validation requires a data set of compounds with known activities for a specific biological target. For the entire set of compounds, all differences are computed between the metric values and between the biological activities. Each metric difference and the corresponding activity difference is graphed in a scatter plot.

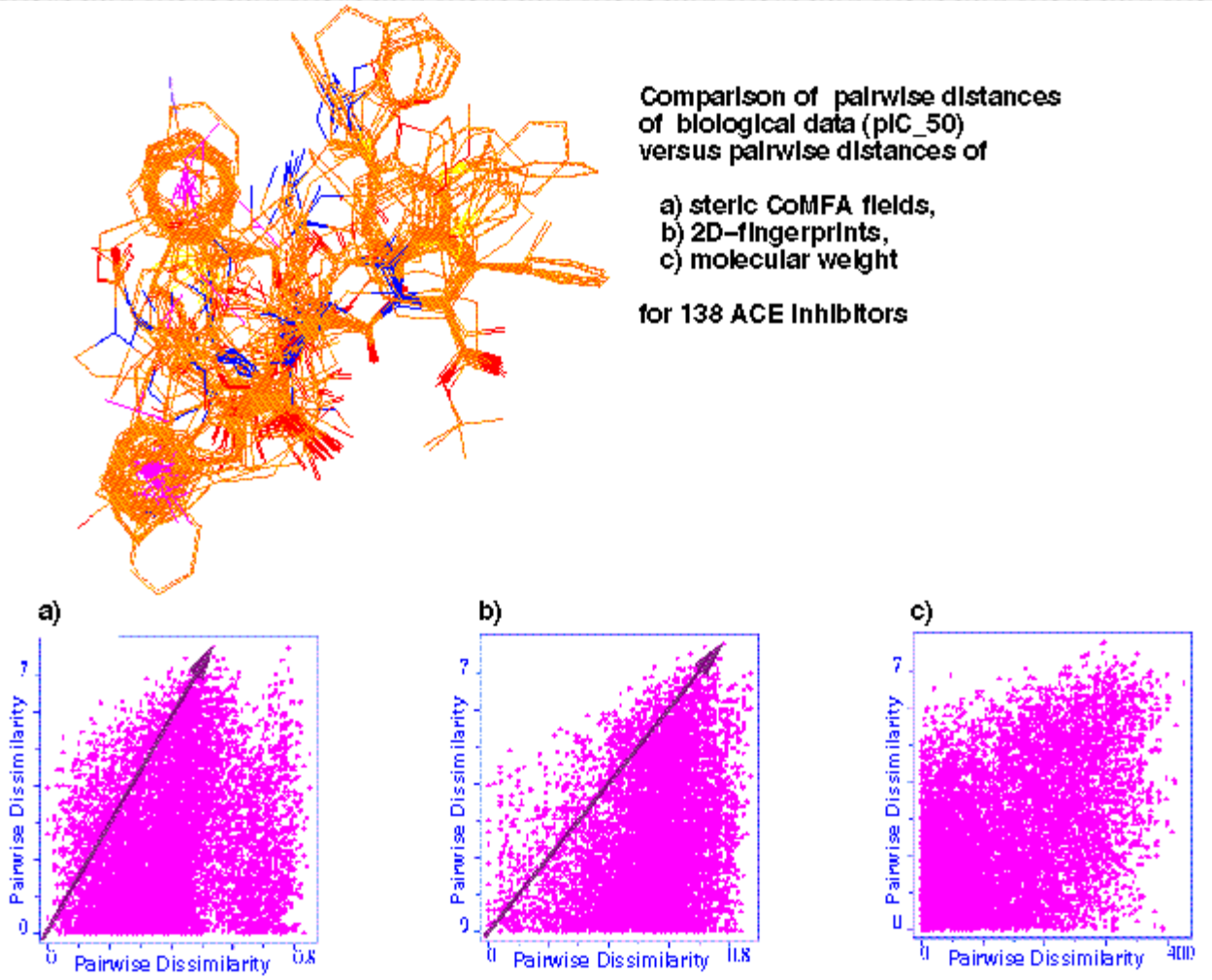



*Fig. 2: Comparison of pairwise distances of biological data (pIC_50 data (7)) with various molecular metrics for a set of 138 Angiotensin-Converting Enzyme (ACE) inhibitors (8), with 9453 data points [n*(n-1)/2]. The molecular geometries and the superposition rule (molecules displayed in the upper left part) used to derive the molecular steric field similarities correspond to literature data (9). The plots on the bottom part contain the pairwise biological activity on the y-axis versus a) CoMFA steric fields, b) 2D-fingerprints and c) Molecular weight as reference descriptor. The maximal slope for the first two molecular metrics is indicated with an arrow in a) and b). Similar graphs for other QSAR series were used to derive the neighborhood radius for molecular steric fields and 2D fingerprints (10).*

If the metric correlates well with biological activity, a characteristic pattern emerges. A valid metric will always produce a void region in the upper left of the plot. The physical interpretation of this phenomenon is that two compounds with similar metric values (small

differences) will also have similar activities (small differences). This phenomenon is called a neighborhood behavior as compounds in the same neighborhood have the same values (activities). In an invalid metric, the converse will typically be true, i.e. two very similar compounds ( as measured by the invalid metric) may have very different activities( one active and the other inactive).

Structurally similar molecules acting on similar biological targets were used to derive quantitative measures for local diversity. Similarity radii of 0.85 for 2D fingerprints were found to be a good threshold value to distinguish between similar and dissimilar compounds, when correlating structural information with biological activity (11).

2D fingerprints represent each molecule by a string of 0's and 1's in a linear bitmap. Each bit in the fingerprint is related to the presence or absence of particular fragments in the molecule. The similarity between two fingerprints a and b can be described by the Tanimoto coefficient T(a,b), which is the number of bits set to 1 in both fingerprints to the number of bits set to 1 in either, i.e. the number of common substructures devided by the number of substructures which appear in only one of the fingerprints. The dissimilarity is then defined as 1-T(a,b). Fig 3 illustrates the effect of small structural variations on the pairwise Tanimoto coefficient for an example of six substituted benzenes.

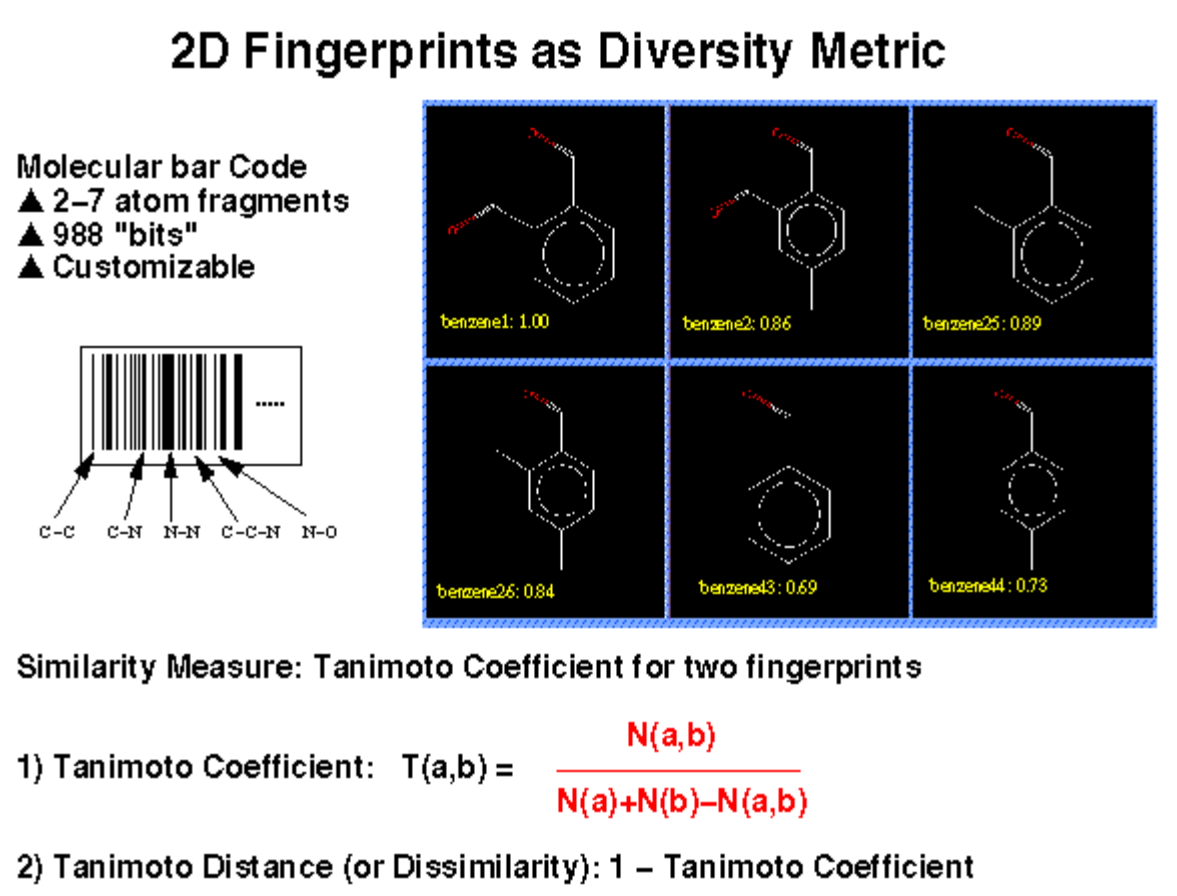



*Fig. 3: Definition of 2D fingerprints as descriptor to quantify molecular diversity and definition of the Tanimoto coefficient to measure 2D similarity or dissimilarity. The inserted panel displays an ensemble of six substituted aromatic compounds and their corresponding Tanimoto coefficient, obtained from pairwise comparison of individual 2D fingerprints with the upper left compound being used as a reference.*

## 2.2. Compound Selections and Database Comparisons

Pairwise Tanimoto dissimilarities have been used to build up diverse samples from the fingerprint data by a maximum dissimilarity method. In this approach the first compound is selected at random from the dataset and then the most dissimilar compound is identified and added to the set of selected compounds. At each successive iteration the unselected candidate which is most dissimilar to those already selected is added to the selection list until a predefined number of compounds has been obtained or no further compound which is more dissimilar than a threshold of the Tanimoto distance, can be found. This method has led to diverse subsets which compare well to sets obtained by the more rigorous approach of hierarchical cluster-based selection.

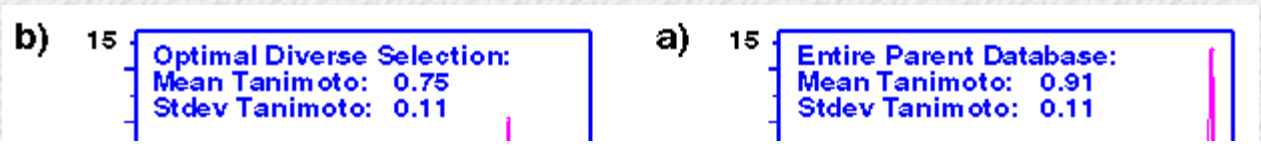
To evaluate the self-similarity of a database and its similarity to other databases (virtual libraries, diversity selections) the following procedure was employed. For each compound in a *reference* database the pairwise Tanimoto coefficient based on 2D fingerprints for the most similar structure in a *candidate* database is computed and a histogram of the similarity index distributions is plotted (Figures 4). Using the same database as *candidate* and *reference* database allows the evaluation of its self-similarity, if identical structures are not considered for the comparison.

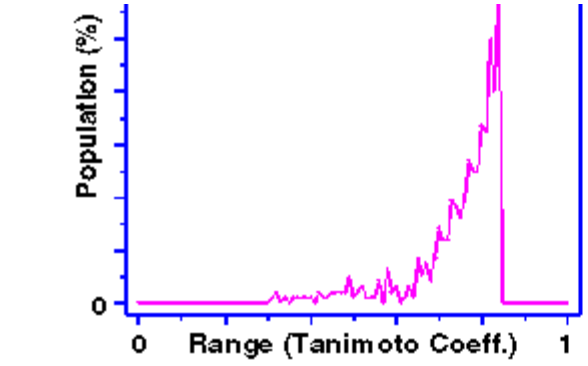
## 3. Results and Discussion

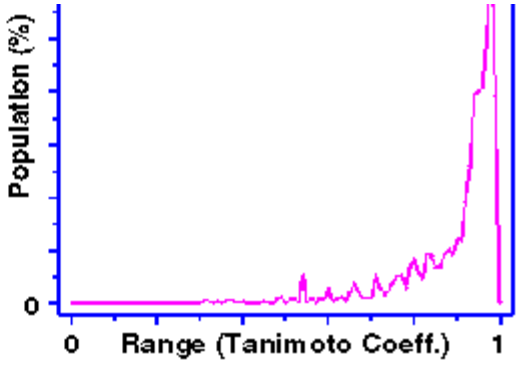
### 3.1. Application to an Existing Commercial Database

For 1283 biologically active structures from IC93 an optimal diverse subset of 487 compounds (38 %) was obtained using 2D fingerprints and a similarity radius of 0.85. The self-similarity plots for the IC93 database (Fig. 4a) and this subset (Fig. 4b) show that the mean Tanimoto coefficient for the subset is remarkably lower than for the entire database (0.75 compared to 0.92, respectively). When computing the similarity between subset and parent database (Fig. 4c), no compounds in the entire IC93 with pairwise Tanimoto coefficients lower than 0.85 were found. In the parent database the mean Tanimoto coefficient is 0.95. Hence, this selection produced a less redundant subset, while not loosing information. Moreover, it can be shown that this optimal diverse subset indeed does cover all 55 biological classes (11), while 10 consecutive random selections of 487 compounds on average did not cover 13.1 % of the classes, thus demonstrating the inappropriateness of a random selection strategy.

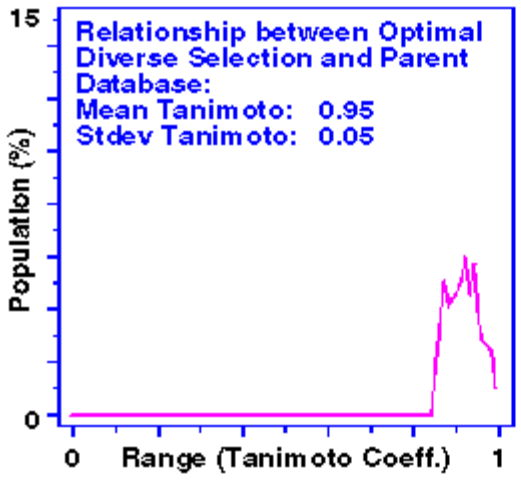
Another interesting question is, whether this subset enhances the diversity of a commercially available database like the MAYBRIDGE catalogue (12). Therefore the most similar compounds within the MAYBRIDGE database to each molecule in the subset were identified and the corresponding Tanimoto coefficient were computed. A mean coefficient of 0.43 (Fig. 4d) suggests that the subset indeed would be a complementary addition to the MAYBRIDGE database.



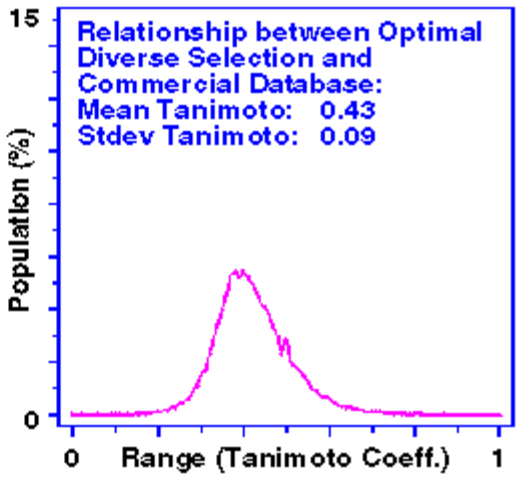

Population (%)
0
0
Range (Tanimoto Coeff.)
1


Population (%)
0
0
Range (Tanimoto Coeff.)
1

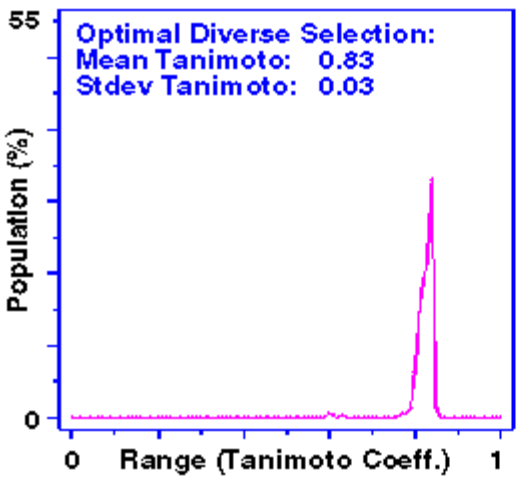
c)


15
Relationship between Optimal Diverse Selection and Parent Database:
Mean Tanimoto: 0.95
Stdev Tanimoto: 0.05
Population (%)
0
0
Range (Tanimoto Coeff.)
1

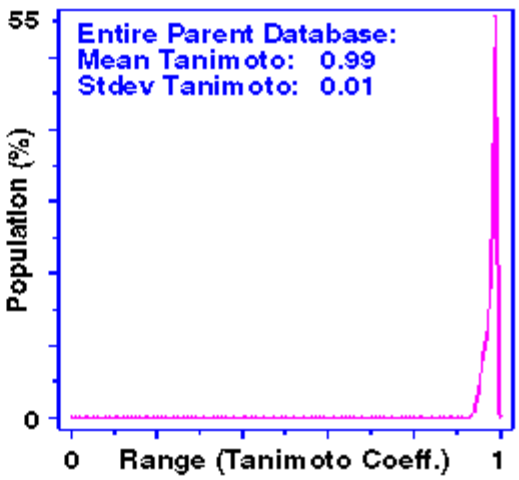
d)


15
Relationship between Optimal Diverse Selection and Commercial Database:
Mean Tanimoto: 0.43
Stdev Tanimoto: 0.09
Population (%)
0
0
Range (Tanimoto Coeff.)
1

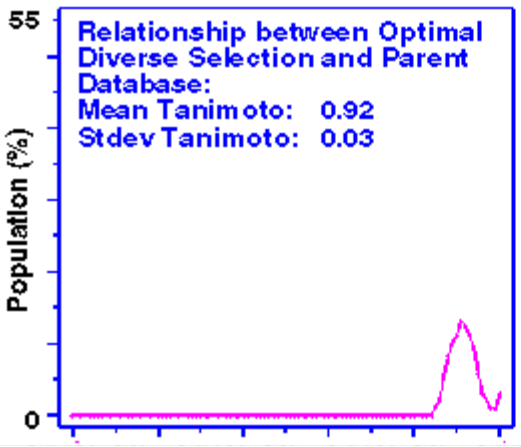
f)


55
Optimal Diverse Selection:
Mean Tanimoto: 0.83
Stdev Tanimoto: 0.03
Population (%)
0
0
Range (Tanimoto Coeff.)
1

e)


55
Entire Parent Database:
Mean Tanimoto: 0.99
Stdev Tanimoto: 0.01
Population (%)
0
0
Range (Tanimoto Coeff.)
1

g)


55
Relationship between Optimal Diverse Selection and Parent Database:
Mean Tanimoto: 0.92
Stdev Tanimoto: 0.03
Population (%)
0

h)

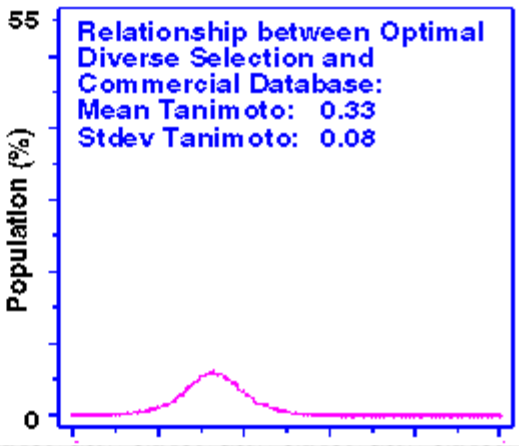

55
Relationship between Optimal Diverse Selection and Commercial Database:
Mean Tanimoto: 0.33
Stdev Tanimoto: 0.08
Population (%)
0


*Fig. 4: Database similarity plots. For each compound of a candidate database the most similar compound of a reference database (excluding the identical compound) is identified and the corresponding Tanimoto coefficient is plotted vs. its population. High mean Tanimoto coefficients indicate a high degree of similarity of the compounds in the dataset. a) Self-similarity of 1283 structures of the IC93 database, b) self-similarity of the optimal diverse subset from IC93 database, c) comparison of the IC93 database with its diverse subset and d) a comparison between the MAYBRIDGE catalogue and the diverse IC93 subset, e) self-similarity of 10752 structures of the virtual combinatorial benzodiazepine database, f) self-similarity of the optimal diverse subset from the benzodiazepine database, g) comparison of the benzodiazepine database with its diverse subset and h) a comparison between the MAYBRIDGE catalogue and the diverse benzodiazepine subset.*

## 3.2. Application to a Virtual Combinatorial Benzodiazepine Library

There are numerous small molecular systems containing large scaffolds that are of current interest in the search for novel biological activity, assuming that the scaffold itself provides a good framework for constructing bioactive compounds. This concept was demonstrated in some studies of benzodiazepines (4,13) as potential b-turn mimetics in peptides and proteins showing interesting biological properties in several biological systems. While such scaffolds are useful for the development of pharmacophoric hypotheses and lead refinement, only relatively small variations are made to an unchanging core, thus leading to highly similar databases.

To demonstrate this, a virtual library of 10752 benzodiazepines was generated. The SYBYL module LEGION was used to mimic the combinatorial synthesis in a reaction oriented approach and to create the full product matrix using three sets of reagents. Four different aminobenzophenones were selected as basic scaffolds and were combined with 96 amino acids and 28 alkylhalides, obtained as hitlists from 2D searches in the UNITY version of the National Cancer Institute's database (123,000 compounds) (12).

Subsequently, all compounds which do not violate the similarity radius were selected from this initial library, resulting in an optimal diverse subset of 298 molecules (3 %). The mean Tanimoto coefficient dropped from 0.99 (Fig. 4e) for the parent database to 0.83 (Fig. 4f) for the subset. Again a plot of Tanimoto coefficients vs. population for a comparison between child and parent database demonstrates the absence of holes (Fig. 4g). As before, the most similar compound within the MAYBRIDGE database to each molecule in the benzodiazepine subset was identified and computed the Tanimoto coefficient. The low mean Tanimoto coefficient of 0.33 for 298 structures again clearly demonstrates that this small subset is a good and diverse addition to the MAYBRIDGE database (Fig. 4h).

This study shows that only 3 % of the synthetically possible benzodiazepine library is sufficiently dissimilar to be used within a lead finding program in the pharmaceutical industry, while all other 97 % of the compounds (i.e. 97 % of the costs) would only produce redundant chemical structures carrying redundant biological information. The result for the IC93 database based on much more scaffolds is different: only 62 % of the structures are redundant. As shown the optimal diverse subset does not only minimizes the chemical redundancy, but also maintains the complete coverage of all 55 biological classes, in contrast to simple random selections.

## 4. Conclusion

Following the paradigm that structurally similar molecules exhibit similar biological activities a strategy to design optimal diverse compound libraries has been applied. A representative subset for a given structure database has been determined without loss of chemical information, but with minimal redundancy. Furthermore it could be shown that still all biological classes were found to be represented in this optimal diverse selection.

The strategy uses molecular structures, properties and metrics to identify those molecules likely to display biological activity, but sufficiently different to minimize the redundancy problem. This should have a tremendous impact on the efficiency of synthesizing and testing compounds, thus dramatically lowering the costs associated with such a project.

---

## References:


1. M. Johnson, G.M. Maggiora, Concepts and Applications of Molecular Similarity, New York, Wiley, 1990.
2. Molecular Similarity in Drug Design, P.M. Dean (Ed.), Chapman and Hall, London 1995.
3. see for example T. Carell, E.A. Wintner, A.J. Sutherland, J. Rebek Jr., Y.M. Dunayevskiy, P. Vouros, Chemistry & Biology 1995, 2, 171-183.
4. (a) B.A. Bunin, J.A. Ellman, J. Am. Chem. Soc. 1992, 114, 10997-10998, (b) B.A. Bunin, M.J. Plunkett, J.A. Ellman, Proc. Natl. Acad. Sci. U.S.A. 1994, 91, 4708-4712.
5. SYBYL Molecular Modelling Package, Version 6.2 and 6.22, Tripos Inc., 1699 S. Hanley Road, St. Louis, MO 63144, USA.
6. UNITY Chemical Information Software, Version 2.5, Tripos Inc., 1699 S. Hanley Road, St. Louis, MO 63144, USA.
7. S.A. DePriest, D. Mayer, C.B. Naylor, G.R. Marshall, J. Am. Chem. Soc. 1993, 115, 5372-5384.
8. see for example: D.W. Cushman, M.A. Ondetti, Hypertension 1991, 17, 589-592.
9. D. Mayer, Naylor, C.B., I. Motoc, G.R. Marshall, J. Comp. Aided Mol. Design 1987, 1, 3-16.
10. H. Matter, J. Med. Chem.1997, 40 (8), 1219-29
11. H. Matter, D. Lassen, Chimica Oggi/Chemistry Today 1996, June 1996, 9-15
12. Maybridge 1995 and NCI 1995 databases are both available from Tripos Inc., 1699 S. Hanley Road, St. Louis, MO 63144, USA
13. S. Hobbs DeWitt, J.S. Keily, C.J. Stancovic, M.C. Schroeder, D.M. Reynolds Cody, M.R. Pavia, Proc. Matl. Acad. Sci. U.S.A. 1993, 90, 6906-6913.


---

### Comments

During 1-30 September 1997, all comments on this poster should be sent by e-mail to ecsoc@listserv.arizona.edu with **F0002** as the message subject of your e-mail. After the conference, please send all the comments and reprints requests to the author(s).

---