J. Med. Chem. **1998,** *41,* 478−488

Random or Rational Design? Evaluation of Diverse Compound Subsets from Chemical Structure Databases

Thorsten Pötter* and Hans Matter†

BAYER AG, Landwirtschaftszentrum, Pflanzenschutz-Forschung, Geb. 6500, D-40789 Monheim, Germany, and TRIPOS GmbH, Martin-Kollar-Str. 15, D-81829 München, Germany

Received February 10, 1997[⊗]

The performance of rational design to maximize the structural diversity of databases for lead finding and lead refinement was investigated. Rational methods such as maximum dissimilarity methods or hierarchical cluster analysis for designing compound subsets were compared to a random approach to study their efficiency for an enhancement of the diversity of three different databases. All investigations were done based on 2D fingerprints as a validated molecular descriptor. To compare the performance of the rational selection methods to a random approach, we additionally used probability calculations. When using maximum dissimilarity-based selections, a single compound can be a member of different neighborhoods as defined by the similarity threshold value, while in hierarchical clustering each compound is assigned to only a single cluster. Therefore the relationship between the similarity threshold of the maximum diversity selection method and a 2D similarity search threshold was studied. In contrast to hierarchical clustering analysis, maximum dissimilarity selections allow to use a similarity threshold for adding a new compound to an already selected compound list. Reasonable values for this similarity threshold are presented here. More diverse subsets were designed using maximum dissimilarity selections, which cover more biological classes than using random selections. An optimally diverse subset without redundant structures containing only 38% of one original dataset was generated, where no structure is more similar than 0.85 to its nearest neighbor, but all biological classes were represented. When it is acceptable to cover only 90% of all biological targets, 3.5−3.7 times more compounds need to be selected using a random approach than in a rational design approach. Such coverage rate shows the highest efficiency of design techniques compared to a random approach. In those subsets no compound is closer than 0.70 to its nearest neighbor. Furthermore a comparative molecular field analysis (CoMFA) is used to evaluate designed and randomly chosen subsets for a database consisting of inhibitors of the angiotensin-converting enzyme. It was shown that designed subsets using maximum dissimilarity methods lead to more stable quantitative structure−activity relationship (QSAR) models with higher predictive power compared to randomly chosen compounds. This predictive power is especially high when there is no compound in the test dataset with a similarity coefficient less than 0.7 to its nearest neighbor in the training set.

1. Introduction

Today the evaluation of chemical structure databases and the design of compound subsets is important in order to maximize resources for a successful and timely discovery of new interesting compounds. Hence, the identification of redundant compounds based on molecular similarity considerations[1] is a key requirement of today's novel chemistry techniques such as combinatorial organic synthesis[2] and high-throughput screening. Any reduction of the number of compounds to be synthesized and/or tested, while only reducing the amount of redundancy within a database, but not introducing any voids, should have a dramatic impact on research efficiency and costs associated.[3] Useful high-throughput screening projects do not depend only on the size of the library but also on molecular ensembles without redundant information.[4]

The advent of the concept of molecular diversity[5] stimulated many investigations toward diverse compound selections for synthesis and biological testing. This concept is based on the *similar property principle*,[6] which states that structurally similar molecules should exhibit similar physicochemical and biological properties. This further implies the prediction of unknown target properties for a molecule based on known values for similar compounds. It should be possible to select a representative subset of compounds covering the entire property space of a structure database. Recently some work was done by different groups to investigate which physicochemical measure of similarity translates best to biological activity.[7]

In this article, the selection of diverse chemical compound subsets for biological screening based on a computational chemistry approach is investigated in detail. The superiority of rational library design methods today is still a question;[8] therefore a comparison between an efficient design and a random selection is

* To whom all correspondence should be addressed. Tel: ++49-2173-38-3379. Fax: ++49-2173-38-4945. E-mail: thorsten.poetter.tp@bayer-ag.de.

† Current address: Hoechst AG, Computational Chemistry, Core Research Functions, Building G 838, D-65926 Frankfurt am Main, Germany.

⊗ Abstract published in *Advance ACS Abstracts,* December 15, 1997.

S0022-2623(97)00087-3 CCC: $15.00
Published on Web 01/28/1998

of great interest. Hence, maximum dissimilarity and hierarchical clustering methods for designing compound subsets were compared to random selections in order to obtain information about the efficiency enhancement using those rational techniques. All rational compound selections are based on 2D fingerprints as molecular topological descriptors containing information about the presence or absence of molecular fragments. In earlier studies, a similarity radius for 2D fingerprints could be estimated, and compounds within this similarity radius of another molecule were shown to have comparable biological properties.[7] It was not the aim of this study to compare various descriptors or various selection techniques, but to compare some well-accepted and representative techniques to a random selection.

Compound selections to find subsets of diverse chemical structures were done in the present study on a public database containing 1283 compounds active in 55 biological classes with several topologically diverse templates. The question here is, how many compounds must be selected to get a representative from every biological class of this database using random or rational methods? A second test example, a database of 334 compounds containing members from 11 different quantitative struture−activity relationship (QSAR) target series was studied with the same question in mind. This database has the advantage that the biological data are obtained in well-defined biological assay systems under identical conditions and are not compiled from literature data obtained from different laboratories, which could be a potential source of uncertainty for the first dataset. Finally a database of 138 inhibitors of the angiotensin-converting enzyme[9] was investigated. Here designed or randomly chosen subsets were used as a training set for 3D-QSAR studies based on the CoMFA methodology (comparative molecular field analysis). The resulting 3D−QSAR models were subsequently used to predict the biological activities of the remaining compounds not included in the training set. As this database contains only molecules acting on the same target, the question to answer was, how many compounds are needed in order to generate a valid and predictive QSAR model? This study was carried out with the aim to compare the ability of a randomly selected or designed subset to forcast biological activities of related compounds. The corresponding results should reflect how CoMFA operates on diverse or similar training sets and thus provide guidelines on how to design informative series for generating and validating sound and predictive QSAR models.

The choice of 2D fingerprints as descriptors for compound selections is based on previous investigations[7] showing their superiority compared to other 2D or 3D molecular descriptors. In these earlier studies it was found that compound subsets without any compound closer than 0.85 to another one (measured using the Tanimoto coefficient[10] of 2D fingerprints as a dimensionless metric) are able to span the entire biological property space of a database. For all biological targets at least one representative, bioactive compound is sampled. Thus a removal of redundant structures should lead to a child database spanning the same physicochemical diversity space with a smaller number of compounds, which still carry the same information as the parent database.[11] Such a novel subset designed using a validated descriptor should also cover the entire biological property space. We will refer to such a database as an "optimally diverse" database in the following context.

2. Methods

2.1. General Methods. All calculations and modeling work were done using the program package SYBYL, versions 6.22 and 6.25.[12] Database manipulations were carried out using UNITY 2.5 database management tools[13] in connection with the SYBYL module SELECTOR to analyze and compare databases. In the following context, we will refer to the following UNITY programs: *dbdiss*, a program to select compounds based on a maximum dissimilarity method; *dbsearch*, a program for similarity searching in UNITY databases. Detailed descriptions of these programs can be found elsewhere,[13] while a brief explanation of the underlying ideas is given in section 2.3. Automation of many procedures was done using the SYBYL Programming Language (SPL) and UNIX shell scripts.

2.2. Computation of Descriptors. 2D fingerprints were generated using the program UNITY (version 2.5). Those descriptors contain information about the presence of molecular fragments in a binary format. For a given chemical structure, a list of all possible fragments of a particular length is generated. The presence of a specific fragment turns on a bit in this bitstring. Due to the large number of existing fragments in a single molecule, it is not possible to assign one bit to only a single fragment. The fingerprints used in the present study were set up as follows: Bits 1−85 encode the presence of two-atomic fragments without taking hydrogens into account. Bits 86−184 encode non-hydrogen-containing three-atomic fragments; the presence of four-atomic fragments including hydrogens is encoded in bits 185−333. Finally four- to six-membered atomic fragments without hydrogens are encoded in the majority of bits from 334 to 928. In addition to this view on molecular fragments, the presence of characteristic groups, rings, or atom types is encoded in the remaining 60 of the total 988 bits.[14] For those features, multiple occurrences are measured and lead to more predefined neighboring bits set to 1. This way of storing molecular information now allows to quantify the similarity of two molecules based on various similarity coefficients, like the Tanimoto or cosine coefficient. A detailed comparison of both types of similarity descriptors is given in ref 15. Both coefficients are based on the number of bit positions set in both individual bitstrings for both molecules normalized by the number of bits set in common. They differ, however, in the exact way of scaling. The Tanimoto coefficient is widely used nowadays in database analysis, as it has certain properties making the work with larger datasets very efficient. Due to the superior features in terms of speeding up a database comparison and selection, we decided to investigate the Tanimoto coefficient in this study. One limitation of this metric is that the triangle inequality does not hold, while it is fulfilled for the cosine coefficient. A similarity coefficient of 0 means that both structures have no "1" bits in common; there is no intersection between both sets of fragments. In contrast, a value of 1 indicates that the fingerprints are identical. Similar features are present in those molecules as far as the fingerprint descriptor is concerned. Examples of different molecules and the corresponding Tanimoto coefficients can be found elsewhere in the literature.[11] This implementation of 2D fingerprints was chosen, as an earlier study[7c] has revealed a similar performance for different types of hashed fingerprints (e.g., in the implementation of the Tripos UNITY and Daylight software packages).

2.3. Compound Selection and Comparison. Two alternative approaches for compound selections and classifications were used based on the distance matrix (or dissimilarity matrix) between every pair of compounds. The fastest approach to select a representative subset is called the *maximum dissimilarity* method.[16,17] The implementation of this strategy

begins with a random selection of a seed compound. Then every new compound is successively chosen such that it is maximally dissimilar from all members of the previous subset. For this investigation, we use (1 − Tanimoto coefficient) as a dissimilarity measure. This coefficient was computed between every candidate molecule and all members of the already selected subset to identify the next compound to be selected. This entire process will be terminated either when a maximum number of compounds has been selected or when no other molecules can be selected without being too similar to one of the already selected subset members. This latter criterion avoids the selection of redundant compounds. As this method takes a random starting point, the variance in the results was checked by comparing various selections with a preset maximum number of compounds on one of the described databases.

Cluster analysis as an alternative rational technique offers more specific control by assigning every single structure to a group of compounds. Hierarchical clustering[18,19] does not require any prior assumptions about the number of clusters; small clusters with a very close relationship between their members are nested within larger clusters containing more dissimilar structures. Many different clustering methods have been reviewed in the literature,[18] and there are no a priori guidelines, which will be most appropriate for a particular dataset, although some methods show a better performance for grouping similar compounds.[7c] As it was not our intention to compare the performance of different hierarchical clustering methods, we have chosen the hierarchical agglomerative cluster-center method (i.e., the distance between two clusters equals the distance between the two cluster centroids). Initial studies[7d] to other hierarchical clustering methods (single linkage, the distance between the closest pair of data points in both clusters; complete linkage, the distance between the most distant pair of data points in both clusters; average linkage, the average of all pairwise data points between two clusters; have shown only small differences when comparing their ability to separate active from inactive compounds of one of our later described databases. After the clustering process, the structure closest to the center of a cluster is selected as the representative structure.

Random selections were used to generate subsets of identical sizes for comparison based on the C-routine *rand()*. The distribution of active compounds in those subsets will be compared to corresponding subsets generated using rational methods.

The mean Tanimoto coefficient for each compound subset is computed as an average using the Tanimoto coefficients for every structure to its nearest neighbor. This similarity index distribution can be used to generate a histogram, and the maximum Tanimoto coefficient as the closest pair of two compounds within the entire dataset is extracted for analysis.

2.4. Probability Calculations. A. Random Selection. Probability calculations were used to compare the performance of rational compound selections to a random approach. A compound with a reported specific biological activity is referred to as a hit. This activity can be measured using an enzyme assay. The probability p to find exactly $n1$ hits using n selections in a database with a total of N compounds and $N1$ hits for this particular target is given by eq 1:

$$p(N1,n1,N,n) = \frac{\binom{N1}{n1} * \binom{N - N1}{n - n1}}{\binom{N}{n}} \quad (1)$$

where $N1$ is total number of hits for a particular target in the database, $n1$ is number of hits to be selected, N is total number of compounds in the database, and n is number of tries to select exactly $n1$ hits.

The probability to find at least one hit for a biological target class (btc) can be calculated as:

$$WP_{\mathrm{btc}} = \sum_{i=1}^{n1} p(N1,n1,N,n) \quad (2)$$

For datasets containing more than one biological target class, the probability to cover all classes by at least one hit per class by selecting n compounds is given by eq 3:

$$P = \prod_{\mathrm{btc}=1}^{M} WP_{\mathrm{btc}} \quad (3)$$

where M is number of target classes.

The question, how many biological target classes are covered by a random selection of n compounds, can now be addressed by computing the mean of all probabilities for all individual targets (eq 4):

$$NP = \overline{WP_{\mathrm{btc}}} \quad (4)$$

B. Selection by Hierarchical Clustering. The success of a hierarchical cluster analysis is given by the proportion in eq 5:

$$\mathrm{cl}_i = \frac{\text{no. of hits in cluster } i}{\text{total no. of compounds in cluster } i} \quad (5)$$

Then the probability to find at least one hit by hierarchical clustering of the dataset and selecting one compound randomly of each cluster is

$$WC_{\mathrm{btc}} = 1 - \prod_{i}^{n} (1 - \mathrm{cl}_i) \quad (6)$$

where n is number of clusters.

The probability to find no hit $(1 - \mathrm{cl}_i)$ is used, because it can be calculated easier than the probability to find exactly one hit, two hits, etc. The probability to cover all targets by at least one hit per biological target class is given in eq 7:

$$P = \prod_{\mathrm{btc}=1}^{M} WC_{\mathrm{btc}} \quad (7)$$

where M is number of target classes.

The question of how many biological target classes are covered by hierarchical clustering and a random selection of one compound from each cluster again is answered by the mean of all individual probabilities:

$$NP = \overline{WC_{\mathrm{btc}}} \quad (8)$$

2.5. Comparative Molecular Field Analysis. CoMFA[20−22] is a useful QSAR technique with numerous known applications. Here this 3D-QSAR technique is used to further evaluate designed and randomly chosen subsets. The starting geometries and superposition rules for the investigated dataset consisting of 138 angiotensin-converting enzyme (ACE) inhibitors were taken from DePriest et al.[9] Following the definition of a superposition rule for the 3D representations of ACE inhibitors, the steric and electrostatic interaction energies between a probe atom and every structure are calculated at the surrounding points of a predefined grid, using a volume-dependent lattice with 2.0-Å grid spacing, a positively charged carbon atom and a distance-dependent dielectric constant. The magnitude of the regions was defined to extend the ensemble of superimposed conformers by 4.0 Å along the principal axes of a Cartesian coordinate system. The maximum field values were truncated to 30 kcal/mol for the steric and ±30 kcal/mol for the electrostatic interaction energies. For points "*inside*" a molecule (determined by a steric energy value of 30 kcal/mol), no electrostatic energy was computed. Those field values were set to the mean of the corresponding column in the PLS

(partial least-squares) analysis. To speed up the analyses and reduce the amount of noise, a column filter was used to exclude the columns with a variance smaller than 2.0 (*minimum σ*). Equal weights were assigned using the CoMFA scaling option.[23] For cross-validation, the leave-one-out method was utilized, as implemented in the program SAMPLS[24] within SYBYL. Unless specifically stated otherwise, default settings for all other parameters in CoMFA were used.

The overall quality of the analyses was expressed by the corresponding cross-validated r^2 value r^2(cv), defined as:

$$r^2(\text{cv}) = (\text{SD} - \text{PRESS})/\text{SD} \quad (9)$$

where SD is the variance of the biological activities of the molecules around the mean values. PRESS represents the sum of the squared differences between the predicted and target property values for every compound. The ideal value of 1.0 is reached when PRESS becomes 0.0 (i.e., the internal prediction is perfect). Hence, this r^2(cv) is considered to be a very critical indicator for the internal consistency of the analysis. The calculation of the predictive r^2 value was based on the molecules in the test set around the mean activity of the training set molecules.

3. Results and Discussion

3.1. Characteristics of the Databases. Three different databases from diverse sources and with different characteristics were investigated. The first database, IC93, is a collection of 1283 biologically active molecules extracted from the IndexChemicus 1993 database.[25] Compounds having similar biological activities were put in similar classes leading to the definition of 77 biological target classes according to the biological activity strings extracted from that database. A detailed listing of all represented biological activities and the population of each class are given in the Supporting Information. However, while some compounds are active in more than one biological class, other classes were only populated by a few members. It should be noted that this database contains for some classes heterogeneous bioactive molecules, which might act on more than one receptor. This is considered to be a potential source of uncertainty. It is likely that a subdivision into more classes corresponding to biological receptors could improve the classification results. However, this database was the only public data collection available for this purpose when this investigation was started.

Initial selections were analyzed using this grouping, while for later investigations, classes with very similar biological activities were grouped together, leading to 55 classes. Details of this grouping are also given in the Supporting Information. As the results for both biological classifications in preliminary studies were very similar, the classification leading to a lower total number of groups and less groups with only a few members was further utilized.

The second database, BAYER, with 334 compounds was generated using various structure−activity series for 11 proprietary and diverse biological targets. Those compounds were put into a single database. One important criterion in the selection of the QSAR series was the different size and degree of similarity within each individual dataset as well as different similarity relationships between different series. In other words, some compounds being active in different series are sometimes quite similar and sometimes very dissimilar.

The third database is a diverse set of 138 ACE inhibitors as first analyzed by DePriest et al.[9] Here quantitative biological activities were available (pIC_{50} values) on a 7-orders-of-magnitude range in a uniform distribution.

To better characterize the structural variance within each individual database and its translation into the 2D fingerprint descriptor, we computed the following statistical data. For the first database, IC93, the average number of bits for all molecules in the 2D fingerprint descriptor file being set to 1 is 19.1% (SD 0.087, highest coverage 42.0%). The average number of atoms in this database including hydrogens is 52.1 (SD 21.6), the average number of bonds 53.7 (SD 21.9), the average number of rings 2.7 (SD 1.4), the average number of rotatable bonds 11.2 (SD 8.1), and the average number of heteroatoms 5.9 (SD 3.7). For the second database, BAYER, we observed an average number of bits set to 1 in the fingerprint file of 22.5% (SD 0.10, highest coverage 28.7%). Here the average number of atoms is 49.2 (SD 12.6), the average number of bonds 51.5 (SD 13.2), the average number of rings 3.1 (SD 0.9), the average number of rotatable bonds 9.6 (SD 3.7), and the average number of heteroatoms 6.9 (SD 2.5). Finally the third database with the 138 ACE inhibitors shows an average number of bits set to 1 in the fingerprint file of 16.6% (SD 0.53, highest coverage 28.7%). For that database the following average values were observed: the average number of atoms 42.0 (SD 16.9), the average number of bonds 42.7 (SD 17.8), the average number of rings 1.7 (SD 1.0), the average number of rotatable bonds 9.4 (SD 4.3), and the average number of heteroatoms 6.7 (SD 2.3). These data show the wide variance within the input 2D fingerprint descriptor datasets and the databases. A detailed analysis of the nearest-neighbor similarities for all three databases to show the degree of redundancy is given as a separate figure in the Supporting Information. The majority of pairwise nearest-neighbor similarities for all three databases shows a Tanimoto coefficient larger that 0.85, while only some structurally unique compounds are present. This degree of redundancy is reflected by high mean Tanimoto coefficients of 0.907 for the IC 93 database, 0.842 for the BAYER dataset, and 0.872 for the ACE138 database. Hence all three databases are interesting candidates to remove redundancy and select diverse subsets.

3.2. Hierarchical Clustering versus Random Selection. First the database BAYER was investigated in detail using hierarchical cluster analysis (agglomerative cluster-center method). After the generation of a dendrogram, the intercluster relationships were used to generate 10, 20, 30, 40, 50, 60, 70, or 80 individual clusters representing the entire database. The separation of active and inactive compounds at different levels in the hierarchical cluster dendrogram is displayed in Figure 1 in comparison to the probability to find at least one active compound using a random selection. This number of selected compounds equals the number of generated clusters, because a single compound is randomly chosen from each individual cluster. The first analysis was done only for one single target in this database out of 11 targets. This should answer the question, how many compounds are needed to find at

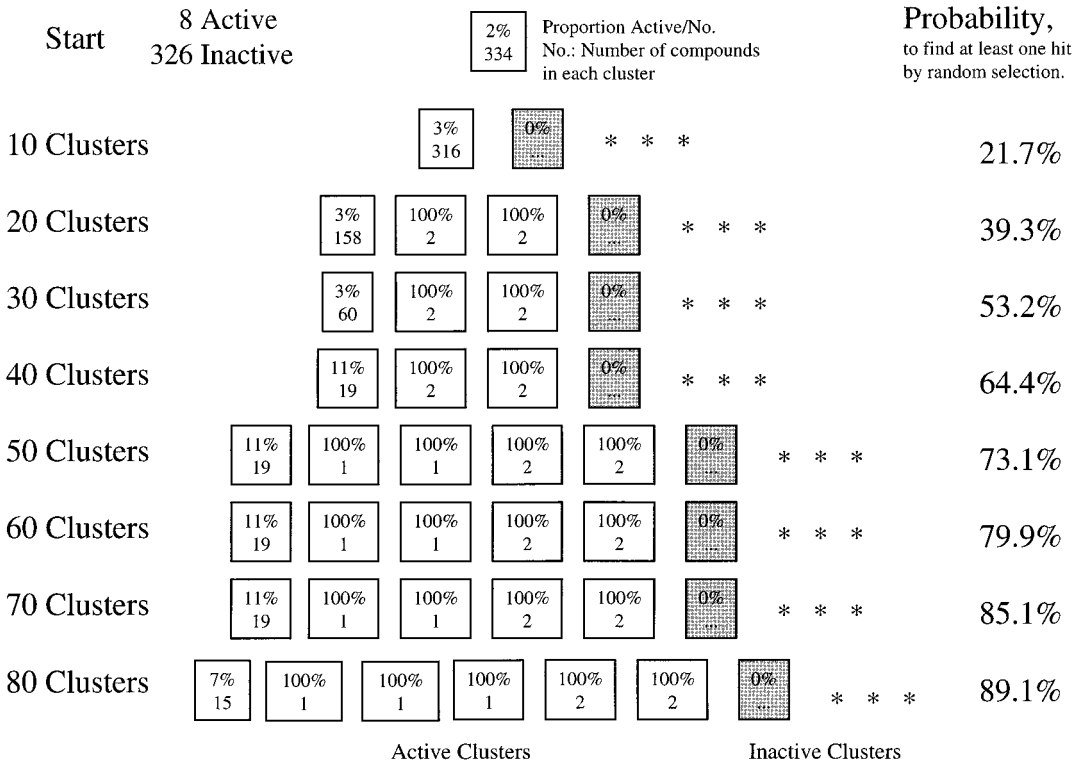

Figure 1. Separation of active and inactive compounds from a hierarchical cluster analysis of the BAYER database given at various levels of the cluster dendrogram (on the left side, 10−80 clusters were generated) in comparison to the probability to find at least on active compound by random selection (given on the right side).

least one active compound out of 8 hits for this single target in all 334 compounds? A cluster containing at least one single biologically active compound is called an "active cluster", and all other clusters are referred to as "inactive clusters". Generating 20 clusters led to three active clusters. In two of them, there are only two compounds, and both are active (Figure 1). Hence a subset of 20 molecules is sufficient to identify at least two compounds showing biological activity for this target. In contrast, the probability to find a single hit for this target using a random selection of 20 compounds is 39.3% (eq 1).

The graph in Figure 2 shows a comparison to cover all biological targets between the probabilities of a random selection and a selection by hierarchical clustering. The individual probabilities were calculated following eqs 3 and 7, respectively.

Using hierarchical clustering a minimum of 40 compounds (12%) of this database must be selected to cover all biological target classes by at least one hit per target. From each cluster, the compound next to the cluster center was used here as the representative compound. In contrast, a random selection of 40 compounds shows that the probability to cover all targets is only ca. 12%. Even a selection of 80 compounds only corresponds to a probability of ca. 65% to cover all targets. This suggests that a random selection is not as efficient as the rational design technique described here.

This retrospective analysis shows that it is possible to represent the structural diversity of a database using a smaller subset without missing interesting biological activity. Other hierarchical clustering methods may even lead to better results. However, one problem remains unsolved: In practice the activity information is not known a priori. The number of clusters to be

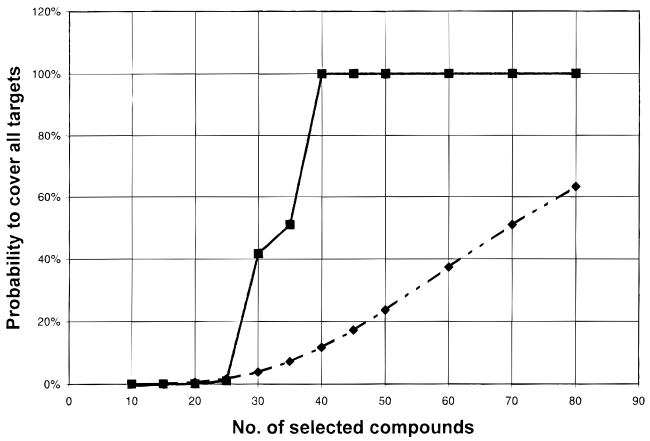

Figure 2. Comparison of the probability to cover all biological targets using a random selection (◆) or a selection based on clusters (■) formed using a hierarchical cluster analysis for the BAYER database with 334 compounds and for a single biological target. The individual numbers were calculated using eqs 3 and 7, given in the text.

generated must be decided based solely on structural information and similarity between molecules. To approach this problem further, the unbiased selection of the optimal number of compounds to choose will be described in the next section.

3.3. Database Representation by Subsets Using Maximum Dissimilarity Methods. A selection of compounds using the maximum dissimilarity method can be terminated when no new molecule can be selected without being too similar to the already selected subset. Hence the similarity of molecules based on the Tanimoto coefficient and 2D fingerprints can be seen as an adjustable parameter of this method. Thus the following calculations were carried out on the IC93 and

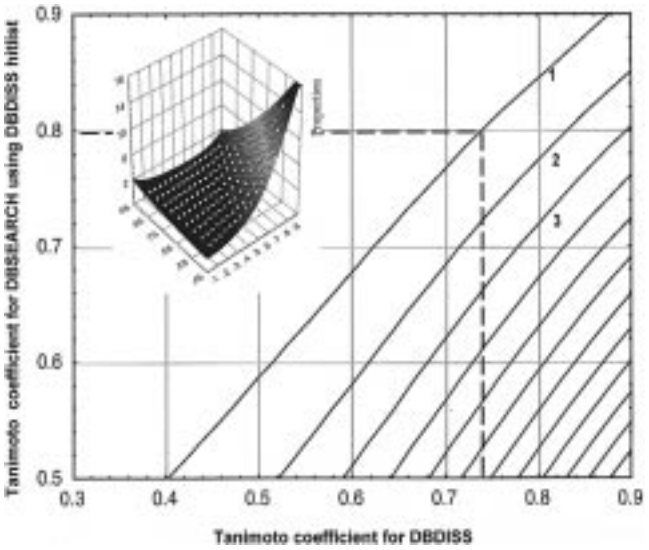

Figure 3. 2D Tanimoto coefficient contour map. The relationship between Tanimoto coefficients used as termination criteria for a maximum dissimilarity selection to generate diverse subsets (*x*-axis) versus the Tanimoto coefficients to extract neighboring compounds in a UNITY 2D similarity search (*y*-axis) is plotted. The proportion used to generate the 2D contour map is defined as the number of similarity search hits divided by the number of compounds in the entire database; this value is computed for every pair of Tanimoto coefficients. Individual contours are shown with a spacing of 1, and for the first three contour lines the corresponding proportions are indicated. The corresponding 3D plot is shown in the upper right corner of the 2D plot for reference only.

BAYER databases to determine, how the entire structural diversity of the database can be described by smaller, but diverse, subsets. For the first step a diverse subset is selected using the maximum dissimilarity method, as implemented in the UNITY program *dbdiss*. Each of the selected compounds is subsequently used as a query in a 2D similarity search using the UNITY program *dbsearch*. This search also uses 2D fingerprints as similarity descriptors and Tanimoto coefficients for structural comparisons. Finally all hitlists from individual searches were combined. For each subset a characteristic proportion defined as total number of hits from the similarity search divided by the number of compounds in the entire database is calculated. In this initial step of this investigation, several diverse subsets were generated with termination Tanimoto coefficients ranging from 0.25 to 0.90 in steps of 0.05. This parameter ensures that the resulting subsets do not contain any pair of molecules that is more similar than this threshold value. For each of the resulting subsets several 2D similarity searches were carried out with a minimum Tanimoto coefficient as the similarity criterion ranging from 0.50 to 0.90 in steps of 0.05.

Both datasets IC93 and BAYER lead to very similar results when analyzing the characteristic proportion in detail. For further discussion the graphs from the analysis of the BAYER dataset are displayed in Figure 3. On the *x*-axis the Tanimoto coefficients used as termination criterion for the maximum dissimilarity selection are plotted versus the Tanimoto coefficients for a 2D similarity search on the *y*-axis. The proportion computed for each *x*,*y*-pair of Tanimoto coefficients is used to generate a 2D contour plot for a detailed analysis. For reference the original 3D plot is shown in the upper right corner of the 2D contour display. Individual contours are shown with a contour spacing of 1; for the first three contour lines the corresponding proportions (dimensionless units) are indicated in this figure. Of course this proportion is high when all compounds from a maximum dissimilarity subset generated with a termination criterion of 0.95 are used for a 2D similarity search with a termination Tanimoto coefficient of 0.5. It was surprising that the optimal proportion of 1 was not obtained using the same Tanimoto coefficients for the diversity selection and similarity search, but using a lower Tanimoto coefficient for the diversity selection. This is graphically shown in Figure 3.

This finding suggests that a particular compound obtained in a maximum dissimilarity subset can be a member of different neighborhoods as defined by the similarity search threshold values. This finding is in contrast to hierarchical clustering methods, where a single compound is always assigned to only one cluster. At the borderline in Figure 3, where this characteristic proportion is equal to 1, the overlaps of different neighborhoods are nearly eliminated and most of the structural diversity is covered. Although this procedure does not prevent from selecting one compound more than once, it shows that the generation of a diverse compound subset, where no structure is more similar than 0.8 to any other member, is possible using a Tanimoto coefficient of 0.74 for the termination of the maximum dissimilarity selection.

The next step of this analysis is now to investigate how similar two compounds must be in oder to take only one for a representative subset. A Tanimoto coefficient of 0.85 was earlier suggested as a threshold value.[7]

3.4. Maximum Dissimilarity Methods versus Random Selections for Global Diversity. For the IC93 database various subsets were selected with an increasing number of members from 60 to 500 compounds in steps of 20 using the maximum dissimilarity method or random selections. To obtain data of higher statistical significance, the results from 100 random selections were averaged for each subset. These data are analyzed in Figure 4.

In Figure 4a the mean and maximum Tanimoto coefficients for all resulting subsets are plotted on the *y*-axis versus the number of compounds in each subset given on the *x*-axis. The mean Tanimoto coefficient is defined as the average value for the Tanimoto coefficients between neighboring pairs, while the maximum Tanimoto coefficient corresponds to the value for the closest pair of compounds within a particular subset. When increasing the number of compounds in a subset, the mean and maximum Tanimoto coefficients are also increased for either the random selection or maximum dissimilarity subset. It can be seen that the maximum dissimilarity method led to more diverse subsets with lower mean and maximum Tanimoto coefficients compared to the randomly selected compounds.

From a detailed inspection of the mean and maximum Tanimoto coefficients for the randomly chosen subset, it is obvious that several redundant structures were sampled in contrast to maximum dissimilarity methods.

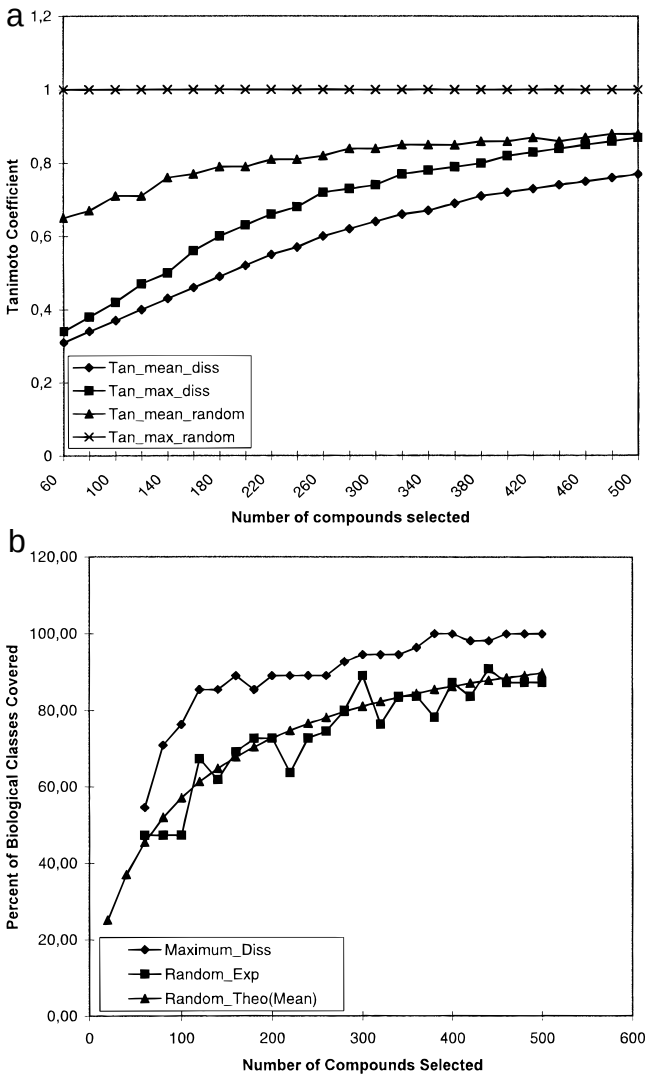

Figure 4. Selection of various compound subsets from the IC93 database using different methods: random or maximum dissimilarity selections (denoted as *random* or *diss*) based on 2D fingerprints. (a) Comparison of the mean and maximum Tanimoto coefficient (denoted as mean or max) for both selections plotted on the *y*-axis versus the number of chosen compounds in a subset given on the *x*-axis. (b) Comparison of the percentage of biological classes covered from the IC93 database plotted on the *y*-axis versus the number of compounds in each subset (*x*-axis) for a maximum dissimilarity selection, an experimental random selection, and the theoretical expectation for a random selection.

This is reflected by the decreased pairwise similarity within those subsets. Hence using 2D fingerprints and maximum dissimilarity methods, more diverse subsets can be generated.

Those subsets also represent more biological classes than the corresponding randomly selected subsets, as can be seen from Figure 4b. The entire biological property space of the original database is better represented by subsets designed using maximum dissimilarity methods. The percentage of represented biological classes for the IC93 database is plotted on the *y*-axis in Figure 4b versus the number of structures within each subset for the random subset (random_exp) and the maximum dissimilarity subset (maximum_diss). In addition, the theoretically expected coverage rates for a random selection are also shown. This latter curve shows great correspondence to the experimental curve for a random selection, where only one of 100 examples is shown. When selecting more than 440 structures, from all biological classes at least one representative is sampled. In contrast, many classes are not represented within the corresponding randomly chosen subsets. A maximum dissimilarity subset with 460 structures, corresponding to a maximum Tanimoto coefficient of 0.85, can be selected without missing any biological information. This maximum Tanimoto coefficient is in good agreement with the similarity radius for 2D fingerprints derived using other methods.[7]

Thus an optimally diverse subset for the 1283 biologically active structures from IC93 is obtained, when selecting 487 compounds (38%) using 2D fingerprints and a similarity radius of 0.85. The mean Tanimoto coefficient for this subset is lower (0.72) than for the entire IC93 database (0.92).

The selection of compounds using a maximum dissimilarity method also has a random component, as its seed structure is randomly picked. Thus the variance in the results was checked by comparing the coverage of biological classes for various selected subsets with 120 as the preset maximum number of compounds on the IC93 database. Although individual structures are different in different subsets, the represented biological classes are similar. For subsets with 460 compounds on the other hand, all biological classes are consistently represented.

These investigations now allow to define an efficiency enhancement for a maximum dissimilarity selection, when compared to a random selection, using the following relationship:

$$\text{efficiency} = N_{(\text{random})}/N_{(\text{maxDiss})} \qquad (10)$$

This answers the question, how many compounds need to be randomly selected ($N_{(\text{random})}$) in order to achieve a comparable coverage rate of biological target classes to the maximum dissimilarity selections ($N_{(\text{maxDiss})}$)? To cover, for example, 90% of all biological classes for the IC93 database, 3.47 times more compounds must be randomly selected compared to maximum dissimilarity methods. Hence the latter method is 3.47 times more efficient (Figure 5). For the IC93 and the BAYER databases this efficiency maximum is identically found at a coverage rate of ca. 90% of all biological targets. For this BAYER database a maximum diversity selection is 3.7 times more efficient than a random selection, as shown in Figure 5. Thus a selection of only 90% of all biological targets is the most efficient selection when comparing those two methods with respect to the two independently compiled datasets studied here. To cover more or less biological target classes using maximum dissimilarity methods, a lower efficiency enhancement is found.

Filling the gap between 90% and 100% coverage rates requires many more compounds. For example more than 440 compounds of the IC93 database are needed to cover 100% of the biological targets instead of about 280 compounds to cover 90%. Thus the efficiency enhancement effect is not longer so dominant. Trying to cover less biological targets than ca. 90% on the other hand becomes in essence almost similar to a random selection.

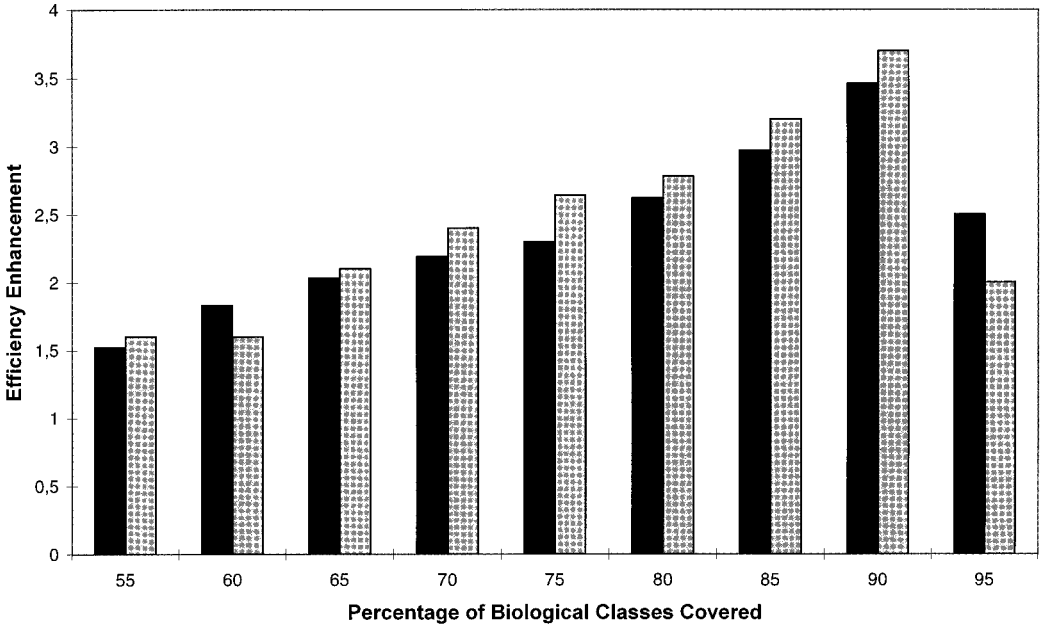

Figure 5. Efficiency enhancement for the IC93 database (black) and the BAYER database (gray) for coverage rates of the biological target classes, computed using eq 10.

We think that allowing some biological targets to be missed when designing new compound subsets is advantageous in terms of efficiency and resource management. This corresponds to an improvement of the efficiency of the screening process, while this selection produces a 10% chance of not getting any hit for a target. Thus it strongly depends on a cost/benefit analysis for a specific screening experiment, whether this procedure is acceptable or dangerous. The aim of this investigation was to provide guidelines for such an analysis.

A similarity radius of 0.85 can be derived for a maximum dissimilarity subset, which covers 100% of the biological targets, while for subsets covering only 90% of the biological classes, a similarity radius (i.e., maximum Tanimoto coefficient) of 0.7 can be found (see Figure 4). This leads to the conclusion that for the design of a compound collection suitable for secondary screening or lead refinement, a similarity radius of 0.85 seems to be suitable in order not to miss important information, while for an initial screening library for a lead discovery program a similarity radius of 0.7 is sufficient, given the enhanced efficiency compared to a random selection and the 90% coverage of biological targets.

3.5. Maximum Dissimilarity Methods versus Random Selections for Local Diversity. Finally the effectiveness of maximum dissimilarity methods as experimental design technique to generate different sets of diverse compounds was investigated. Those compounds were subsequently analyzed using CoMFA as a 3D-QSAR technique.[20−23] The purpose of this study is to illustrate a strategy based on our previous results for the development of a predictive 3D-QSAR model[26,27] using CoMFA. The entire strategy is again based on the assumption that smaller sets of representative compounds, if properly chosen, represent all other compounds in a structurally homogeneous class—a similar assumption to that previously used to design compound subsets for biological screening. This investigation reflects, how CoMFA operates on diverse or similar training sets and thus provides guidelines on how to design informative series for generating and validating sound and predictive QSAR models.

A dataset of 138 ACE inhibitors covering a biological activity of 7 orders of magnitude with a known superposition rule was used to select different compound subsets from 10 to 134 members. Again maximum dissimilarity methods or random selections were utilized to obtain representative subsets. For every subset the mean and maximum Tanimoto coefficients are plotted in Figure 6a versus the corresponding number of compounds within each subset. Again it can be seen that all designed compound subsets are more diverse than any randomly selected subset.

For every compound subset the following strategy was applied to evaluate the predictivity of the QSAR model: A cross-validated PLS analysis was done using steric and electrostatic molecular fields to obtain the cross-validated r^2 value and the optimal number of components. Then a PLS analysis without cross-validation with the optimum number of components was used to compute a conventional r^2 value. Based on this model for various training sets, the biological activities for all other compounds were predicted and the predictive r^2 value was computed following eq 9 for this model. This value clearly indicates the agreement between the test dataset and the 3D-QSAR model. Finally the obtained three different r^2 values for each dataset were plotted against the number of compounds in the training datasets (cf. Figures 6b,c).

From Figure 6b, it can be seen that cross-validated r^2 values are very low when the number of compounds used to generate the QSAR model is lower than ca. 40 molecules. Although the non-cross-validated r^2 value is high for those small subsets, the predictive ability is low as indicated by the small predictive r^2 values. When analyzing subsets with more than 40 structures, the cross-validated r^2 value reaches values between 0.5 and 0.6, while the non-cross-validated and predictive r^2

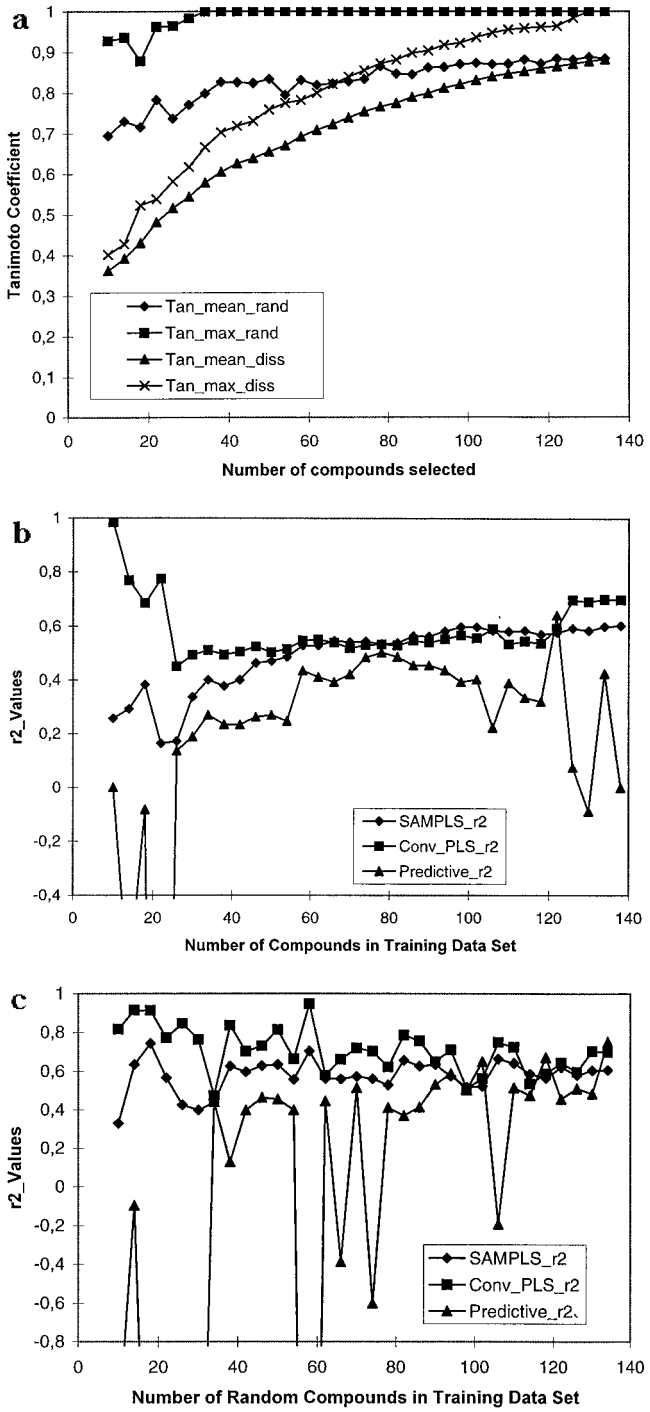

Figure 6. Comparative molecular field analysis of designed and randomly chosen compound subsets for the dataset of 138 structurally diverse ACE inhibitors.[9] (a) Comparison of the mean and maximum Tanimoto coefficients (denoted as mean or max) for random and maximum dissimilarity selections (denoted as *random* or *diss*) plotted on the *y*-axis versus the number of chosen compounds in a subset on the *x*-axis. (b) Comparison of three different r^2 values (cross-validated using SAMPLS, conventional and predictive r^2 plotted on the *y*-axis) for each dataset obtained using maximum dissimilarity methods versus the number of compounds in the training datasets (*x*-axis). (c) Comparison of three different r^2 values (cross-validated using SAMPLS, conventional and predictive r^2 plotted on the *y*-axis) for each dataset obtained using random selection versus the number of compounds in the training datasets (*x*-axis).

values are also acceptable. Now the derivation of the 3D-QSAR model and the prediction of the training dataset are highly reliable. The predictive r^2 value drops significantly when there are more than 130 molecules in the training set, because the corresponding test dataset is statistically not large enough for meaningful analyses. Hence the last values could not be included to derive a trend. Using maximally diverse subsets containing more than 40 molecules led to stable PLS-derived 3D-QSAR models. For every compound to be predicted there is at least a close neighbor in the training set.

To complement this study, another investigation was done with randomly chosen subsets. For each subset, the same type of analysis was applied, leading to the graph displayed in Figure 6c. There is a much higher fluctuation in the cross-validated and conventional r^2 values. However the most dominant feature is the almost complete loss of the predictive ability of individual QSAR models, revealed by very large fluctuations in the predictive r^2. This can be rationalized by the fact that the training datasets are not designed using a rational method. Hence, not only will every prediction be an interpolation, but also a large degree of extrapolation is present. When training subsets are randomly chosen, there is a high risk that variances of several spatial regions are not represented in the training dataset. Hence, a test compound with a substituent in this region cannot be properly predicted by interpolation. For a randomly designed subset that chance for every compound to be predicted to have a close neighbor in the training dataset is low.

Taking only 40 compounds designed using maximum dissimilarity methods led to a maximum Tanimoto coefficient for this dataset of ca. 0.5, while a maximum Tanimoto coefficient of 0.7 is observed with 80 compounds. This latter subset led to better predictive r^2 values and acceptable cross-validated r^2 values. Here no major difference can be seen when taking more designed compounds into account, corresponding to a similarity radius of 0.85. As there are no longer enough compounds in the test dataset, the prediction becomes statistically unstable and the predictive r^2 value is no longer relevant. It can be concluded that designed subsets of compounds using maximum dissimilarity methods lead to more stable QSAR models with higher predictive power. This predictive power is especially high when there are no compounds in the test dataset having a Tanimoto coefficient less than 0.7.

This study clearly shows that 3D-QSAR techniques such as CoMFA will work better, in terms of more stable and reliable statistical results, when a design technique was used to generate a well-balanced training set. It is possible to obtain PLS models of similar quality with a much lower number of compounds using design techniques. Furthermore the predictive power of CoMFA is especially high, if there is a high degree of structural similarity between compounds in the training set and the test set. Thus a careful design of training and test datasets should avoid some synthetic effort and allow to prioritize candidate molecules in a more reliable way.

4. Conclusion

The design of combinatorial libraries or the selection of nonredundant compounds from databases is an

essential problem in the lead-finding process. Several techniques and methods have been suggested to address this need. However, one fundamental question stimulated our research described within this publication: Are designed compound subsets superior in the sense of sampling more biological targets compared to randomly chosen subsets? This question was answered using statistical analyses of several random selections in comparison with different rational design approaches. Based on initial studies, maximum dissimilarity methods and hierarchical clustering techniques were chosen to design compound subsets in a rational manner. As the molecular descriptor, 2D fingerprints were utilized because they were recently shown to be appropriate in selecting representative subsets of bioactive compounds, when comparing the sampling properties of other metrics carrying 2D or 3D molecular information. For our analysis it was important to investigate whether a particular method was able to select group molecules according to their biological properties. Indeed this analysis reveals that a randomly selected compound subset generally represents fewer biological classes than any descriptor-based rational selection.

All three databases used in this study were interesting candidates to lower redundancy and design diverse subsets because of a high degree of very similar structures found by analyzing pairwise Tanimoto coefficients between neighboring pairs. Using hierarchical clustering a subset of 12% of the compounds from the BAYER database was found to represent all biological target classes by at least one hit per target, while the probability to cover those targets using a random selection is indeed very low (only 12%).

Then the relationship between maximum diversity and similarity searches was studied to find a pair of similarity coefficients, which resulted in the ideal proportion of 1 comparing the number of hits from both applications. This optimal proportion of 1 was using a somewhat smaller Tanimoto coefficient for the diversity selection than for the similarity search. This finding suggests that a particular compound obtained in a maximum dissimilarity subset is present in different neighborhoods defined using the similarity search thresholds, which contrasts with hierarchical clustering methods, where a single compound is always assigned to only a single cluster. Using a proportion of 1, the overlap of different neighborhoods is nearly eliminated and most of the structural diversity is covered.

Furthermore allowing some biological targets to be missed when designing new compound subsets is advantageous in terms of efficiency and resource management. It was possible to get information about the efficiency enhancement, i.e., the relationship between the number of compounds to be chosen randomly or using design techniques in order to cover a specific percentage of biological targets. Filling the gap between 90% and 100% coverage rates requires many more compounds, and the efficiency enhancement effect is not longer so dominant. The highest efficiency was found for two unrelated databases at a value of 90% coverage of the biological targets, which translates in both cases to a Tanimoto coefficient of 0.7 for a similarity search. In contrast, a similarity radius of 0.85 can be derived for a maximum dissimilarity subset, which covers 100% of the biological targets. Hence we conclude that for designing a compound collection for secondary screening or lead refinement, a similarity radius of 0.85 seems to be sufficient, while for an initial screening library a similarity radius of 0.7 should be used based on the derived enhanced efficiency.

Hence this detailed retrospective analysis suggests the possibility to represent the structural diversity of a database using smaller subsets generated using rational design techniques without missing interesting biological activities. Thus a more detailed picture of molecular diversity and design techniques begins to emerge and will help to better understand the fundamental relationship between the degree of structural variation and its influence on biological activity.

Acknowledgment. The comments of Dr. R. Cramer III, D. Patterson, Dr. A. Ferguson, Dr. P. Hecht, Dr. R. Clark (Tripos), Dr. M. Schindler, and Dr. F. Reichel (BAYER) are gratefully acknowledged. The authors thank Dr. S. DePriest for making the structures of the ACE dataset available.

Supporting Information Available: Figure with pairwise similarity histograms computed for each compound in all three databases investigated here to its next neighbor based on 2D fingerprints and a table with an overview of the biological classes from the IC93 database (6 pages). Ordering information is given on any current masthead page.

References

(1) *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Chapman and Hall: London, 1995.

(2) For recent reviews, see: (a) Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries. *J. Med. Chem.* **1994**, *37*, 1233−1251. (b) Gordon, E. M.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gallop, M. A. Applications of Combinatorial Technologies to Drug Discovery. 2. Combinatorial Organic Synthesis, Library Screening Strategies, and Future Directions. *J. Med. Chem.* **1994**, *37*, 1385−1399. (c) Madden, D.; Krchnak, V.; Lebl, M. Synthetic Combinatorial Libraries: Views on Techniques and Their Applications. *Persp. Drug Discovery Des.* **1995**, *2*, 269−285. (d) Ellman, J. A. Design, Synthesis and Evaluation of Small-Molecule Libraries. *Acc. Chem. Res.* **1996**, *29*, 132−143. (e) Gordon, E. M.; Gallop, M. A.; Patel, D. V. Strategy and Tactics in Combinatorial Organic Synthesis. Application to Drug Discovery. *Acc. Chem. Res.* **1996**, *29*, 144−154.

(3) Ferguson, A. M.; Patterson, D. E.; Garr, C.; Underiner, T. Designing Chemical Libraries for Lead Discovery. *J. Biomol. Screen.* **1996**, *1*, 65−73.

(4) Moos, W. H.; Green, G. D.; Pavia, M. R. Recent Advances in the Generation of Molecular Diversity. *Annu. Rep. Med. Chem.* **1993**, *28*, 315−324.

(5) 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.

(6) Maggiora, G. M.; Johnson, M. A. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990; pp 99−117.

(7) (a) Patterson, D. E.; Cramer, R. D., III; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of Molecular Diversity Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059. (b) Brown, R. D.; Bures, M. G.; Martin, Y. C. Similarity and Cluster Analysis Applied to Molecular Diversity. Presented at the American Chemical Society Meeting, Anaheim, CA, 1995; COMP3. (c) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584. (d) Matter, H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of 2D and 3D Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.

(8) Young, S. S.; Farmen, M.; Rusinko, A., III. Random Versus Rational. Which is Better for General Compound Screening? *Network Sci.* (electronic publication) **1996**, 2.

(9) DePriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. 3D-QSAR of Angiotensin-Converting Enzyme and Thermolysin Inhibitors: A Comparison of CoMFA Models Based on Deduced and Experimentally Determined Active Site Geometries. *J. Am. Chem. Soc.* **1993**, *115*, 5372−5384.
(10) (a) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Letchworth: Research Studies Press, 1987. (b) Willett, P.; Winterman, V. A. Comparison of Some Measures for the Determination of Intermolecular Structural Similarity. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18−25.
(11) Matter, H.; Lassen, D. Compound Libraries for Lead Discovery. *Chim. Oggi* **1996**, *14* (6), 9−15.
(12) SYBYL Molecular Modelling Package, versions 6.22 and 6.25; Tripos Inc., 1699 S. Hanley Rd., St. Louis, MO 63144.
(13) UNITY Chemical Information Software, version 2.5; Tripos Inc., 1699 S. Hanley Rd., St. Louis, MO 63144.
(14) For details and setup files to compute fingerprints, see: *UNITY Chemical Information Software, version 2.5, Reference Guide*; Tripos Inc.: St. Louis, MO; pp 45−58.
(15) Holliday, J. D.; Ranade, S. S.; Willett, P. A Fast Algorithm for Selecting Sets of Dissimilar Molecules from Large Chemical Databases. *QSAR* **1996**, *14*, 501−506.
(16) Lajiness, M.; Johnson, M. A.; Maggiora, G. M. Implementing Drug Screening Programs using Molecular Similarity Methods. In *QSAR: Quantitative Structure−Activity Relationships in Drug Design*; Fauchere, J. L., Ed.; Alan R. Liss Inc.: New York, 1989; pp 173−176.
(17) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59−67.
(18) Barnard, J. M.; Downs, G. M. Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644−649.
(19) *SYBYL 6.2, Ligand-Based Design Manual;* Tripos, Inc.: St. Louis, MO, 1995; pp 246−255 and references therein.
(20) Cramer, R. D., III; Patterson, D. E.; Bunce, J. E. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(21) Clark, M.; Cramer, R. D., III; Jones, D. M.; Patterson, D. E.; Simeroth, P. E. Comparative Molecular Field Analysis (CoMFA). 2. Towards Its Use with 3D-Structural Databases. *Tetrahedron Comput. Methodol,* **1990**, *3*, 47−59.
(22) *3D-QSAR in Drug Design. Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, The Netherlands, 1993. This includes many applications and cross-references of the CoMFA methodology in medicinal chemistry.
(23) (a) Thibaut, U.; Folkers, G.; Klebe, G.; Kubinyi, H.; Merz, A.; Rognan, D. Recommendations for CoMFA Studies and 3D QSAR Publications. In *3D QSAR in Drug Design. Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, The Netherlands, 1993; pp 711−717. (b) Folkers, G.; Merz, A.; Rognan, D. CoMFA: Scope and Limitations. *Ibid.* pp 583−616. (c) Cramer, R. D., III; DePriest, S. A.; Patterson, D. E.; Hecht, P. The Developing Practice of Comparative Molecular Field Analysis. *Ibid.* pp 443−485.
(24) Sheridan, R. P.; Nachbar, R. B.; Bush, B. L. Extending the Trend Vector: The Trend Matrix and Sample-Based Partial Least Squares. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 323−340.
(25) Index Chemical Database − subset from 1993; Institute for Scientific Information, Inc. (ISI), 3501 Market St., Philadelphia, PA.
(26) Carpignano, R.; Dolci, M.; Scarfi, D. Design of More Informative Molecules for QSAR Study in a Series of Juvenile Hormones. In *Trends in QSAR Molecular Modelling 92, Proceedings of the European Symposium on Structure−Activity Relationships;* Wermuth, C.-G., Ed.; ESCOM: Leiden, The Netherlands, 1993.
(27) (a) Norinder, U.; Hoegberg, T. PLS-Based Quantitative Structure−Activity Relationship for Substituted Benzamides of Clebopride Type. Application of Experimental Design in Drug Research. *Acta Chem. Scand.* **1992**, *46*, 363−366. (b) Norinder, U. Experimental Design-Based Quantitative Structure−Toxicity Relationship of Some Local Anesthetics Using the PLS Method.. *J. Appl. Toxicol.* **1992**, *12*, 143−147. (c) Norinder, U. An Experimental Design Based Quantitative Structure−Activity Relationship Study on β-Adrenergic Blocking Agents Using PLS. *Drug. Des. Discovery* **1991**, *8*, 127−136. (d) Norinder, U. Experimental Design Based 3D QSAR Analysis of Steroid−Protein Interaction: Application to Human CBG Complexes. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 381−389.

JM9700878