Clustering and Rule-Based Classifications of Chemical Structures Evaluated in the Biological Activity Space

Ansgar Schuffenhauer,* Nathan Brown, Peter Ertl, Jeremy L. Jenkins, Paul Selzer, and Jacques Hamon

Novartis Institutes for BioMedical Research, CH-4002 Basel, Switzerland

Received September 14, 2006

Classification methods for data sets of molecules according to their chemical structure were evaluated for their biological relevance, including rule-based, scaffold-oriented classification methods and clustering based on molecular descriptors. Three data sets resulting from uniformly determined in vitro biological profiling experiments were classified according to their chemical structures, and the results were compared in a Pareto analysis with the number of classes and their average spread in the profile space as two concurrent objectives which were to be minimized. It has been found that no classification method is overall superior to all other studied methods, but there is a general trend that rule-based, scaffold-oriented methods are the better choice if classes with homogeneous biological activity are required, but a large number of clusters can be tolerated. On the other hand, clustering based on chemical fingerprints is superior if fewer and larger classes are required, and some loss of homogeneity in biological activity can be accepted.

INTRODUCTION

Partitioning sets of chemical objects into sets of structurally related clusters is an important task in several stages of drug discovery. There are two reasons for this. First, patents aim to cover structural classes rather than cherry-picked individual compounds, and thus, structural classes often coincide with common patent protection. Second, structurally similar compounds are known to also have similar biological activity as it is stated by the similarity-property principle.[1,2] Recently, data from in vitro biological profiling has become available in which a set of compounds is screened against a constant panel of assays in such a way that a complete structure-target IC_{50} matrix is created[3] (www.cerep.fr). Using this data, Fliri et al.[4] showed that compounds exhibiting a similar biological profile are often also structurally similar. Conversely, Barbosa and Horvath [5] state that this is not generally true, since a common absence of biological activity does not require any structural similarity; even a common activity can result from different modes of interaction between the ligand and target. If structures are however detected as similar by a biologically meaningful chemical descriptor, one should expect them also to have a similar biological activity.

For this reason, classification by chemical structure should yield classes with greater homogeneity in terms of biological activity when compared with random partitioning of the data set, as has been exemplified by Böcker et al.[6,7] In this paper, we evaluate the extent to which different classification methods fulfill this expectation on the basis of in vitro bioactivity profile data. Including different structural representations in the study, one can simultaneously evaluate the potential of these representations for the identification of similar biological activity in the absence of a common scaffold ("scaffold hopping")[8,9] by comparing classification results obtained with these representations with those obtained from classification by scaffold. However, it is recognized that, for practical selection purposes, medicinal chemists may still prefer to sample two distinct scaffolds even if they can be assumed to have similar biological activity, because of different synthetic accessibility or intellectual patent coverage issues. However, in case not all scaffolds can be included in screening or followed up, it will still be worthwhile to use bioactivity to select series having the potential to cover a wide range of activity profiles.

One can distinguish two types of classification methods for chemical structures. The first method uses a descriptor vector as a representation of the chemical structure. Binary descriptor vectors are also called fingerprints. These vector descriptors can then be subjected to classification methods that are well-established in multivariate statistics such as cell-based partitioning or clustering.[10] Often, these methods involve some stochastic element, which means that different runs with different random seeds or input sequences of the structures could yield different results. In this type of classification method, the assignment of each structure to a class is dependent on the whole data set, and the further addition of structures to this set can alter the classification of structures already within the set. If the classification is based on a descriptor derived from a systematic enumeration of structural elements, as for example chemical fingerprints are, then there is no a priori knowledge included in the classification, except the generic descriptor calculation rules. Dictionary-based fingerprints such as the MDL keys[11] can potentially contain prior knowledge encoded in the dictionary of structural fragments.

The second type of classification methods is rule-based and relies on expert knowledge to define rules for the determination of structural features defining a chemical class.

* Corresponding author e-mail: ansgar.schuffenhauer@novartis.com.

10.1021/ci6004004 CCC: $37.00
Published on Web 02/08/2007

In drug discovery, the most widespread such rule-based system is classification by molecular frameworks (sometimes also referred to as "Murcko-scaffolds"), which are obtained when all terminal side chains are removed from a chemical structure.[12,13] Such methods have the advantage that they are deterministic and data-set-independent, with the latter property sometimes being referred to as "crispness". This is not only an advantage in terms of computation speed, with a linear scaling with the data set size, but also keeps the classification intact when two data sets are merged. In addition, each class is represented by a common, more or less chemically intuitive, substructural feature representing its "chemotype".

The comparison of the classification algorithms is done as follows: With quantitative, uniformly measured biological profiling data, one can define the biological profile distance between two compounds *i* and *j* on the basis of the profile vector of all assays **a**:

$$D_{i,j} = \sqrt{\sum_{\mathbf{a}}^{\text{assays}} [\Delta \log K_i(i, j, \mathbf{a})]^2} = \sqrt{\sum_{\mathbf{a}}^{\text{assays}} [\Delta \log IC_{50}(i, j, \mathbf{a})]^2} \quad (1)$$

Note that when compared as differences on the logarithmic scale, the K_i values can be substituted by IC_{50} values since, according to Cheng and Prusoff,[14] the K_i is essentially the IC_{50} value multiplied by a factor that depends only on the assay conditions, which on the logarithmic scale is a constant increment being neutralized by forming the difference between two log IC_{50} values measured under the same assay conditions. The smaller a $D_{i,j}$ value is, the more similar is the biological activity profile of the compounds *i* and *j*.

According to Kelley et al.,[15] a classification solution can be judged by two objectives, one being the class spread: the average distance of all compounds in a class, averaged again over all classes. While Kelley et al. based the calculation of the class spreads on the descriptor used in the clustering procedure, we use the distance in the biological profile space instead to calculate the cluster spread since this is the external objective against which we wish to benchmark. Accordingly, the spread sp_k of class k with n_k members is defined as

$$sp_k = \frac{2}{n_k(n_k - 1)} \sum_{i=1}^{i \le n_k} \sum_{j=1}^{i<j} D(i, j) \quad (2)$$

Lowering the average of the spread sp_k in the biological profile space over all classes is the first objective, and lowering the number of partitions is the second. Both objectives are clearly competing since, in the extreme situation, where each member of the data set is assigned to its own class, the lowest possible average spread is achieved, whereas the completely unpartitioned data set is also optimal in terms of the second objective but has the worst possible spread. Kelley et al. combined both objectives in a single function after applying normalization and gave both objectives the same weight. If one wishes to avoid the prior assignment of weights to multiple competing objectives, one can use Pareto analysis[16,17] to compare several solutions. In a Pareto analysis, solutions are ranked according to their dominance in all of the objectives as described in Figure 1a. One solution dominates another if it is superior in all objectives: in our case, if it produces less classes and has a smaller overall spread than another solution. In the case of two solutions where one produces less classes and the other has a lower overall class spread, these solutions have the same rank. The set of nondominated solutions describes the optimal tradeoff surface between the two objectives that may be achieved. Most classification algorithms have parameters which can control the tradeoff between the number of classes and the class spread, and the result of each classification procedure applied to the same data set can therefore be described as a tradeoff surface in the Pareto space (Figure 1b).

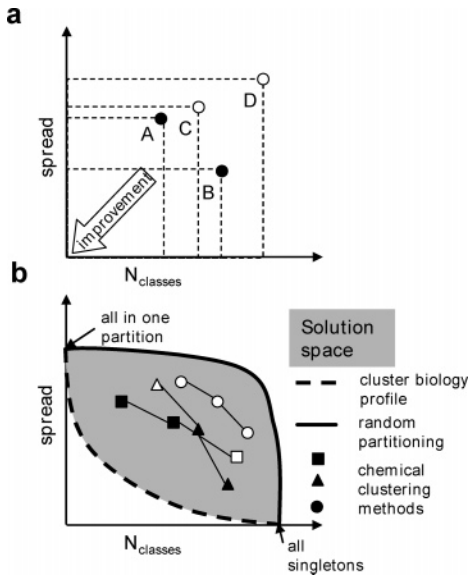

Figure 1. (a) Application of Pareto analysis to classification solutions. Whenever one solution is superior to another solution in all objectives, it dominates the other solution. A and B are nondominated solutions, whereas solution A dominates solution C, and all solutions A, B, and C dominate solution D. (b) Qualitative illustration describing the Pareto analysis for chemical structure-based partitioning in the profile space. The solution space is limited by the tradeoff curves obtained for clustering using the activity profile itself (dashed line) as the optimal solution and random partition of the data (solid line). Chemical classification solutions are expected to be found within this space. If a classification method such as that represented by the circles is consistently dominated by the solutions obtained with another method, it would be on the whole inferior to this other method. On the other hand, the tradeoff curves for two classification methods (such as that represented by squares and triangles) may intersect, in which case it cannot be claimed that one method is generally superior to the other.

As we evaluate the class spread in the biology profile space, it is expected that the optimal solution can be achieved when the profile data is itself used as a descriptor to cluster the molecules. This information is however only available retrospectively and cannot be used for structure-based predictions, although it can serve as an ideal upper-bound comparator. A more realistic upper-bound comparator taking into account the error in biological profile measurement can be estimated by adding normally distributed random noise

with a standard deviation that corresponds to the experimental error of the pIC_{50} determination [typically, log(3)] to the experimental profiling data for the purpose of clustering, but evaluation of the cluster spread criterion is performed with the original biological profiles. On the other hand, chemical structure-based classification solutions are expected to be superior to random partitions of the data set, and hence, the random partitions serve as the lower benchmark. Any tradeoff surface for a chemical structure-based classification is expected to lie in between these two comparator curves.

It is important to consider that the Kelley function was originally devised to select within a cluster hierarchy tree the solution which offers the best tradeoff between cluster spread and number of clusters. We are interested in comparing two alternative classification methods with each other, and therefore, it is to be expected that different classification methods will produce different class size distributions, and in particular a different number of singletons (classes containing only one member). Therefore, a metric that includes all classes is desirable also to take into account the singletons in the calculation of both objectives instead of omitting them completely as is the case with the original Kelley function. Their spread value was set to zero. If this is done, the simple, nonweighted average over all the class spreads sp_k favors unduly methods which split off a larger number of singletons and retain otherwise some large classes as opposed to methods that produce more equally sized classes. In the extreme case, one arbitrary data point removed from a large data set to form a singleton will reduce the nonweighted average spread by half, since the singleton has a spread of zero and the spread of the remaining class will be almost the same one as that of the whole set. To avoid this scenario, we decided to use a weighted average spread SP in our Pareto analysis where the class spreads sp_k are weighted according to the size of the cluster n_k according to the following equation, with n_{classes} being the total number of classes and N the number of structures in the data set:

$$\mathrm{SP} = \sum_{k=1}^{n_{\text{classes}}} \frac{n_k}{N} \mathrm{sp}_k \qquad (3)$$

With this, we have a method that compares how well alternative classification methods fulfill the objective of producing classes with compounds that have a similar biological profile. When two classification methods produce solutions fulfilling the objectives equally well, it is of interest if they are classifying the data in a similar way. For this purpose, the adjusted Rand index as described by Hubert and Arabie[18] can be used: it describes the probability that a pair of objects is either in both classification solutions in the same class or in both solutions in a different class, adjusted by the coincidence of classification which can be expected to occur randomly. The adjusted Rand index R is calculated on the basis of the contingency matrix of two classification solutions with the elements n_{ij} as follows:

$$R = \frac{\sum_{i,j}\binom{n_{ij}}{2} - E}{0.5\left[\sum_i \binom{n_{i\bullet}}{2} + \sum_j \binom{n_{\bullet j}}{2}\right] - E}$$

$$\text{with } E = \frac{\sum_i \binom{n_{i\bullet}}{2} \sum_j \binom{n_{\bullet j}}{2}}{\binom{N}{2}}$$

$$\text{and } n_{i\bullet} = \sum_j n_{ij}$$

$$\text{and } n_{\bullet j} = \sum_i n_{ij} \qquad (4)$$

In cases where $R = 1$, the classification solutions are identical, whereas in the case of $R = 0$, there is only as much overlap between the classes as there can be expected to occur randomly.

For the clustering methods, we used several descriptors, including those that are both conformation-independent and -dependent. The conformation-independent descriptors included chemical fingerprints such as in the UNITY fingerprints (www.tripos.com), which are based on linear connection paths, and the Pipeline Pilot FCFP_4 fingerprints (www.scitegic.com) based on circular substructures which have shown to be highly effective in chemical similarity searching.[19] In addition, we used numerical descriptors embodied as Similog keys,[20] based on an occurrence count vector of graph-distance-based pharmacophore triplets, as well as a vector of computed physical chemical properties comprising molecular weight; log*P*; polar surface area; and the numbers of rotatable bonds, hydrogen bond donors, and acceptors. As conformation-dependent descriptors (3D descriptors) we used feature point pharmacophores (FEPOPS),[21] in which each structure is reduced to a central four-point pharmacophore, and radial distribution functions (RDFs)[22] encoding atomic charge and the interatom distance distribution. As in the original literature described for the FEPOPS descriptors, the seven most diverse conformers are used, whereas the RDF codes are based on a single conformation as generated by the Corina[23] program.

Numerous clustering methods were also applied. The divisive *K*-means (DivKM) clustering procedure[7,24] is used throughout and is known to be very efficient on large data sets. In addition, other clustering or partitioning techniques were used in combination with some of the descriptors to cover methods that are frequently used in the use-case scenarios of the individual descriptors. Self-organizing maps[22,25] (SOMs) were used in combination with the RDF descriptors. The clustering method implemented in the Pipeline Pilot software, based on the OptiSim algorithm[26] (PPClust), is used in combination with both FCFP_4 and UNITY fingerprints as well as FEPOPS. Since a cell-based partitioning founded on the principal components (PCs) of physical chemical properties has been popular in the selection of compounds,[27−29] we included this as well in combination with the physical chemical properties. Cell-based partitioning using principal components (PCA_CELL) also has the advantage that it is deterministic, although still data-set-

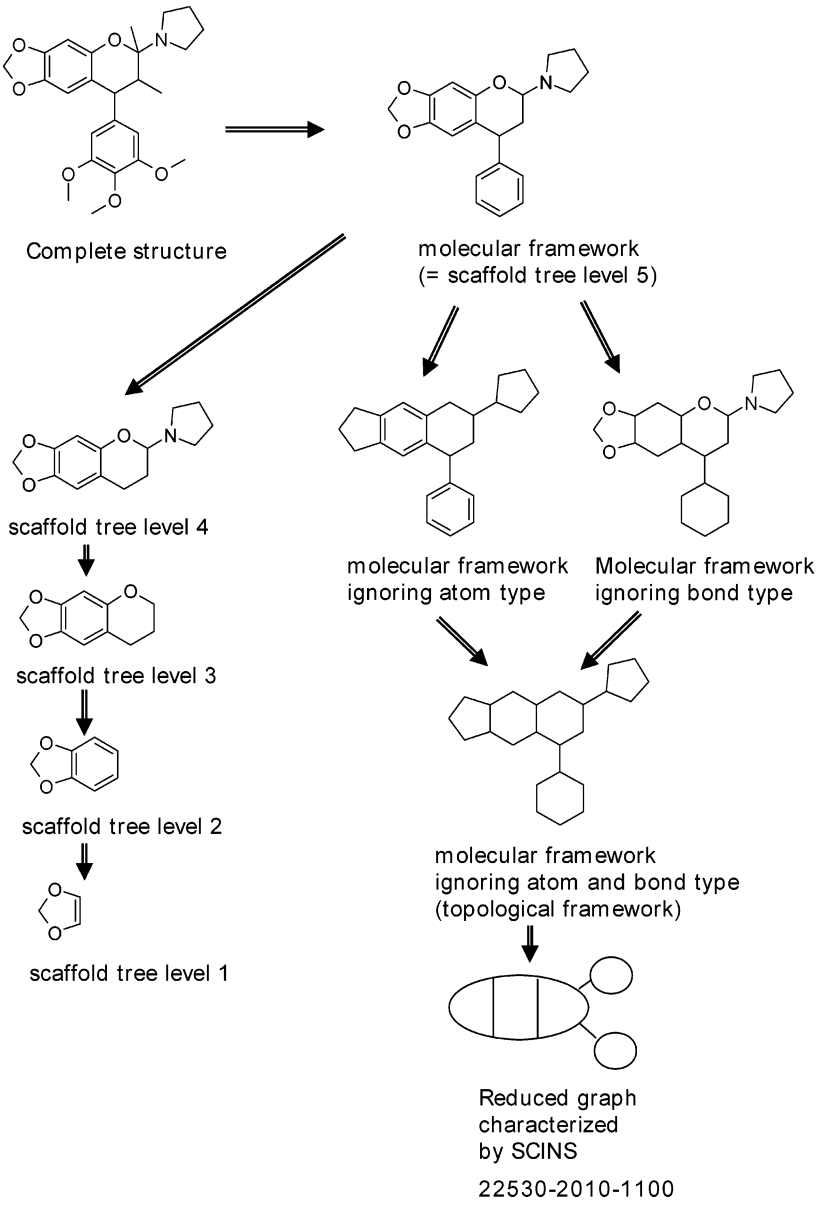

Figure 2. Abstractions of a structure out of the NCI cancer database. Both abstractions shown rely on the molecular framework obtained by pruning all terminal side chains. The scaffold is further generalized by removing atom- and/or bond-type information and, finally, by describing it through the SCINS code ("framework hierarchy"). The Scaffold Tree method generalizes the scaffold by the iterative removal of rings according to prioritization rules. The most prominent rule is to reduce the number of linkage bonds, and therefore, the two linked monocycles are removed first. A second rule with lower priority aims to retain a maximum number of heteroatoms, and consequently, the *n*-heterocycle is removed after the phenyl ring, and in the level 2 scaffold, the dioxolane ring is retained instead of the pyrane ring.

dependent. As rule-based classification, we grouped the compounds according to their molecular framework,[12] obtained by removing all terminal side chains, as well as the frameworks derived from these when either the atom-type information or the bond-type information or both together are discarded.[30] Further abstraction can be reached by grouping the scaffolds according to their scaffold identification and naming system (SCINS) code, which contains only the main features of each scaffold (see the appendix). This cascade of abstractions will be referred to as the "framework hierarchy".

As an alternative way to group scaffolds, we used a hierarchical classification of scaffolds based on the prioritization of the rings contained in them (Scaffold Tree). In this procedure, rings are removed iteratively from each molecular framework according to prioritization rules, and the scaffolds leading to the same remaining subscaffold are grouped together.[31] The main principle (consistent with the way of thinking of chemists) is to remove peripheral, linked, or heteroatom-deficient rings first and retain central, heteroatom-rich, fused rings as a "parent core". This has the effect that scaffolds which share the same core but have different cyclic substituents can be recognized as related scaffolds, which would not be possible using the flat partitioning according to the substituent-pruned scaffold. This method is an enhanced version of an approach used for the hierarchical classification of natural products published by Koch et al.[32] but does not use the information about the frequency of occurrence of scaffolds in the data set any more as a decision criterion and is therefore no more data-set-dependent. The two different ways of rule-based abstraction are illustrated in Figure 2.

The requirement for having a complete compound IC_{50} value matrix measured on a uniform assay panel without

missing values limits the choice of the data sets. Compilations of high-throughput screening (HTS) data are not suitable for our purpose since inactivity is usually not confirmed in these assays and the choice of which of the primary hits are to be submitted to IC_{50} determination is often already influenced by cherry-picking capacity, meaning that an absent IC_{50} value of a compound that has been submitted to the primary screen does not necessarily mean that the compound is inactive.

We used three data sets for this study, two from the Novartis internal biological profiling and one extracted from the National Cancer Institute (NCI) cancer cell-line screening data. The first Novartis data set contains pharmacology safety profiling data of 1006 compounds on 27 assays, mostly aminergic G-protein coupled receptors (GPCRs) and ion channels. The second data set contains the screening data of 3633 compounds screened on a panel of 20 protein kinases. The third data set extracted from the NCI cancer screening data comprises 7747 compounds screened with growth inhibition values (GI_{50}) from 35 cancer cell lines. In this data set, the missing stereochemical data did not allow the calculation of 3D descriptors.

In order to obtain the final structure−activity data matrixes described here without missing values, we had to remove from the three original data sets compound records with many missing assay results, as well as assay data for assays in which only a few compounds had been screened. For further details about data set preparation, please refer to the methods section.

COMPUTATIONAL METHODS

Data Set Preparation. The NCI cancer data set was downloaded as an SDF file from NCI's Web site (http://cactus.nci.nih.gov/ncidb2/download.html). The original file contained 32 557 structures for which growth inhibition activity data were already given as pGI_{50}. Then, from the screening data of those cell lines, data points were removed for those which had pGI_{50} values for less than 90% of the structures. A total of 36 cell lines remained. Now, all structures were removed which did not have pGI_{50} data for these remaining cell lines, leaving 9066 structures surviving standardization with the Pipeline Pilot software. After molecules with reactive functional groups were filtered out with the dbslnfilter tool by Tripos (www.tripos.com), 7747 structures remained. The pGI_{50} for compounds determined as inactive had already been set to 4 in the original data set. This data set is available as Supporting Information.

The two in-house data sets could be used without structural preprocessing. However, IC_{50} data were converted to pIC_{50} values, thereby setting $pIC_{50} = 4$ for all data where there was no activity found within the dynamic range of the assay. Statistical characteristics of the data sets are given in the Supporting Information.

Descriptor Calculation. Pipeline Pilot software was used to calculate FCFP_4 fingerprints, which were then folded to a fingerprint of 2048 bit length. UNITY fingerprints were calculated with UNITY software by Tripos. Physical chemical properties were calculated by Pipeline Pilot using the Alog*P* model by Ghose−Crippen[33] and the topological polar surface area model by Ertl et al.[34] For Similog, FEPOPS and RDF code calculation in-house programs were used.

Clustering. Divisive *K*-means clustering was performed with the divkm clustering software from Digital Chemistry (www.bci.gb.com or www.digitalchemistry.co.uk) using the Soergel (1-Tanimoto) distance measure between all fingerprint descriptors and the Similog keys, while the Euclidean distance was used for all other descriptors—these being the distance metrics usually used in combination with the respective descriptors. The cluster molecules component in Pipeline Pilot was used for the OptiSim-like clustering (PPClust). Clustering of the biological profiles was done with the *K*-means algorithm implemented in the R-statistics package (http://www.r-project.org/).

For the SOM, an in-house implementation of the Kohonen maps described by Gasteiger and Zupan[25] was used. The principal component analysis of computed physical chemical properties was performed with the SIMCA-P software by Umetrics (www.umetrics.com). For each set, two PCs were retained. The following cell-based partitioning was again done with Pipeline Pilot. The partitioning into cells was conducted such that the outmost cells in each dimension were centered at the values ranking at the top 5% and bottom 5% of all values for this dimension, and the PC space included was dissected in the desired number of cells with an approximately equal length of edges in both PCs.

Clustering with the Multiconformation FEPOPS Descriptors. When using the FEPOPS descriptors for similarity searching, typically seven conformations leading to the most diverse FEPOPS vectors were used, and only the one leading to the highest similarity value is taken into account. For this study, we needed an unambiguous assignment of a structure to a cluster. However, using several conformers could lead to a situation where different conformers could be members of different clusters. In this situation, a tie-breaking procedure is required. Two procedures were used. In the first, all cluster members were ranked according to their distance to the cluster center and a structure was assigned to the cluster, where the respective conformer as a cluster member had the lowest distance rank. This method will be referred to as "dist". The second method used a voting procedure assigning a structure to the cluster having a majority of its conformers as members. This tie-breaking procedure will be referred to as "maj". If one procedure failed to resolve a tie between two clusters, then in each case the other was used as a secondary tie-breaking criterion.

Rule-Based Classification. The determination of molecular frameworks and their further abstraction were performed within PipelinePilot as well as the calculation of the SCINS codes. The Scaffold Tree[31] classification by iterative removal of the rings was performed with a proprietary program based on the Molinspiration toolkit (www.molinspiration.com). Because of the high fraction of natural products in the NCI cancer data set, the structures of this had been in silico degylcosidated before computing the scaffold tree in order to avoid the tree being dominated by the glycoside rings.[32]

Repetition of Clustering Runs. All nondeterministic partitioning methods were applied five times using each time a different, pseudo-randomly permuted sequence of the input records to ensure a different initialization even for such programs in which the random seed could not be controlled. Random grouping was performed by assigning each record to a partition based on pseudo-random numbers. The permutation of the experimental profiles with random noise was conducted by generating normal distributed random numbers with a standard deviation of log(3) using the rnorm

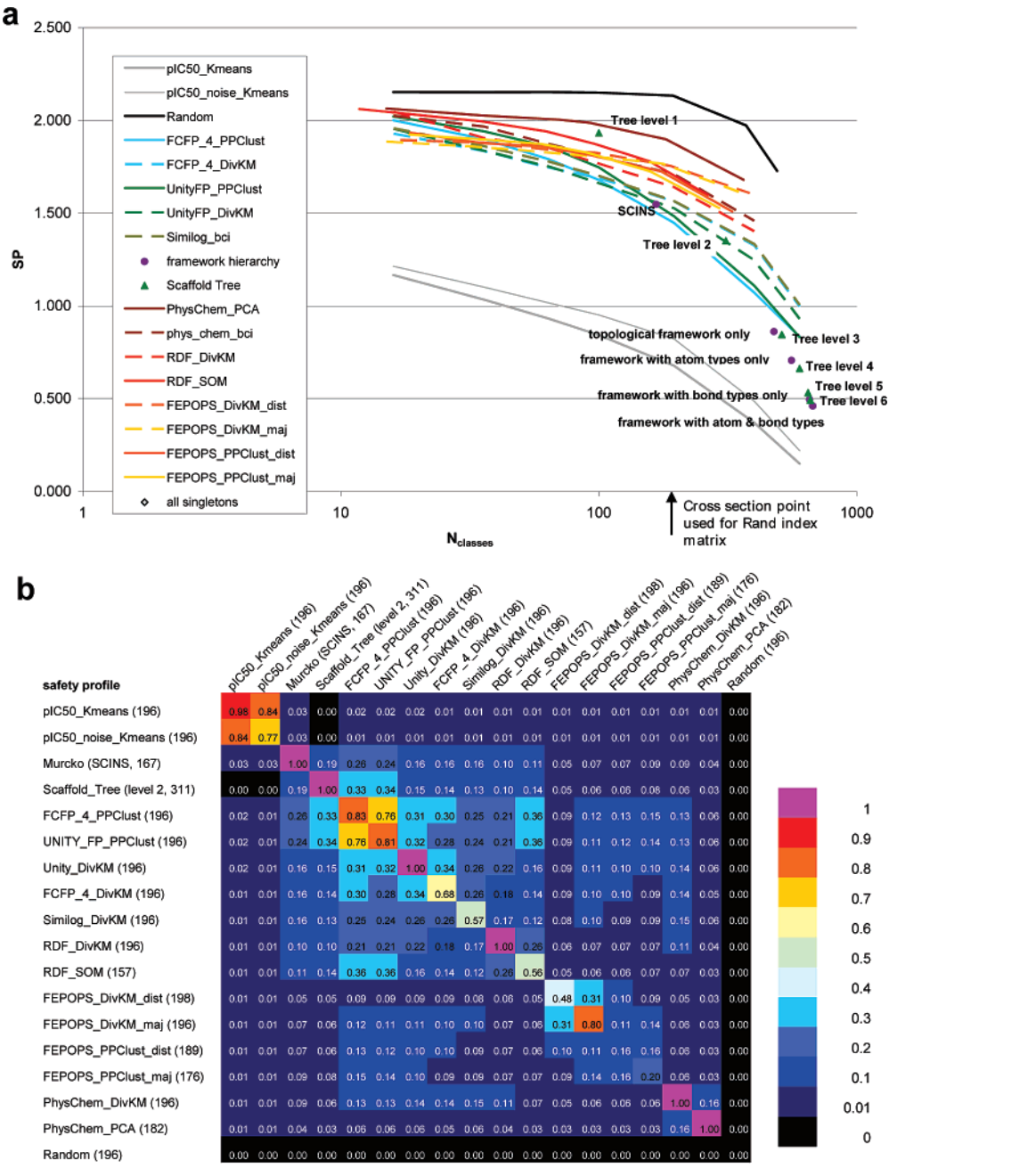

Figure 3. (a) Pareto analysis plot of the safety pharmacology profiling data set. In the *x* axis, the number of classes is displayed on a logarithmic scale. The *y* axis is the average class spread according to eq 3. (b) Matrix of the adjusted Rand indices comparing the classifications obtained with the different methods, averaged over the different runs. The diagonal elements show the average adjusted Rand index according to eq 4 between different runs of the same method. The numbers of classes used are shown in parentheses following each method.

function in the R statistics package. These random numbers were added to all pIC_{50} values which were above the threshold of $pIC_{50} = 4$. If this operation leads to a pIC_{50} value below the threshold of 4, then the pIC_{50} value is set to 4, and out of the pIC_{50} having the threshold value of 4, one was picked randomly and set to the original pIC_{50} value of the permuted data point. This maintained the original threshold, the fraction of the values above it, and the mean pIC_{50}.

RESULTS

Pareto Analysis. The Pareto plots for each of the data sets showing the tradeoff between the number of classes and the spread of the class in the activity space is shown in Figures 3−5. In all three data sets, the tradeoff curves obtained for the different structure-based classification methods are between the boundary curves obtained by clustering with the biological activity data and the random partitioning. The standard deviation of the cluster spreads, obtained with the five runs for each of the nondeterministic methods, was generally rather small and rarely exceeded 0.01. The deviation of the tradeoff curve obtained with the perturbed pIC_{50} data from the curve obtained with the original pIC_{50} data was always larger than the statistical errors of the clustering procedures themselves.

In all three data sets, the intersection between the tradeoff curves of different methods was observed. The Pareto frontier forming the nondominated solutions in the objective space

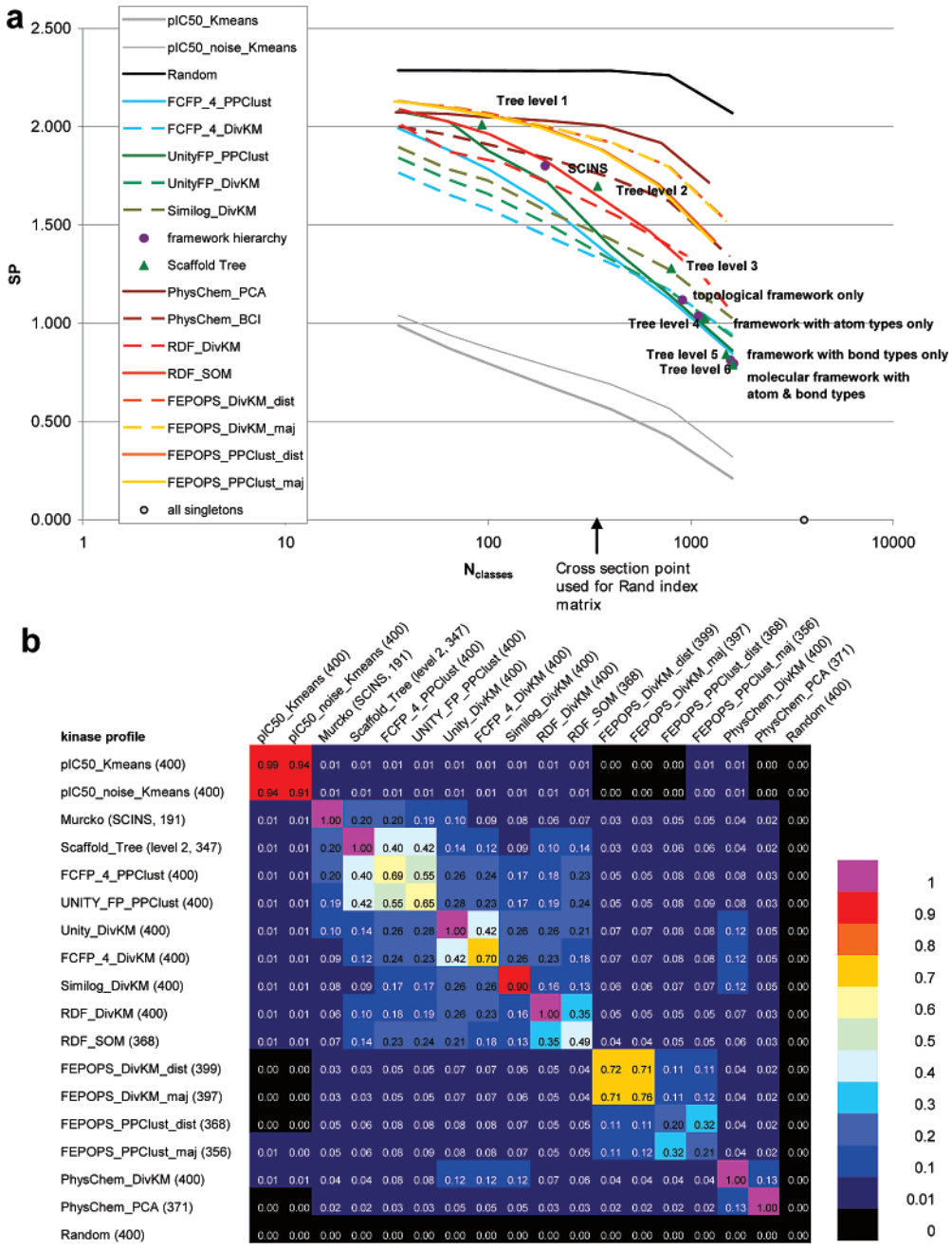

Figure 4. (a) Pareto analysis plot and (b) matrix of adjusted Rand indices for the kinase profile data subset. For details, see the caption of Figure 3.

for the structural clustering is formed in different sections by different classification methods. This means that one overall optimal classification method cannot be identified. In general, the PPClust method of Pipeline Pilot, especially when combined with the FCFP_4 or UNITY fingerprints, is the best solution, if the goal is to have a larger number of partitions with on average five or less members. In the region of smaller numbers of classes, with on average more members, the Pareto frontier is formed by the tradeoff curves of the divisive *K*-means clustering, combined again with the FCFP_4 or UNITY fingerprints. Clearly less optimal are the classification methods based on physical chemical properties especially combined with cell-based partitioning. Also, the classifications obtained using RDF codes either together with divisive *K*-means or self-organizing maps were biologically less homogeneous than those obtained with the fingerprints.

In the analysis of the rule-based methods, in all cases, the number of classes is very high when using the molecular frameworks without further abstraction, leading to an average number of class members smaller than or equal to two. This shows that all data sets are rather diverse. The further reduction of the scaffolds in the Scaffold Tree method gave the reduction to three to five rings and results which were almost as good as the clustering results with Pipeline Pilot, with the exception of the kinase profile set where the two ring scaffolds did not yield a considerably less optimal classification than the fingerprint-based clustering methods. Reduction of the scaffolds by the tree method to a single ring did not provide a better solution than most of the other structural methods, except the physical chemical property-based partitions.

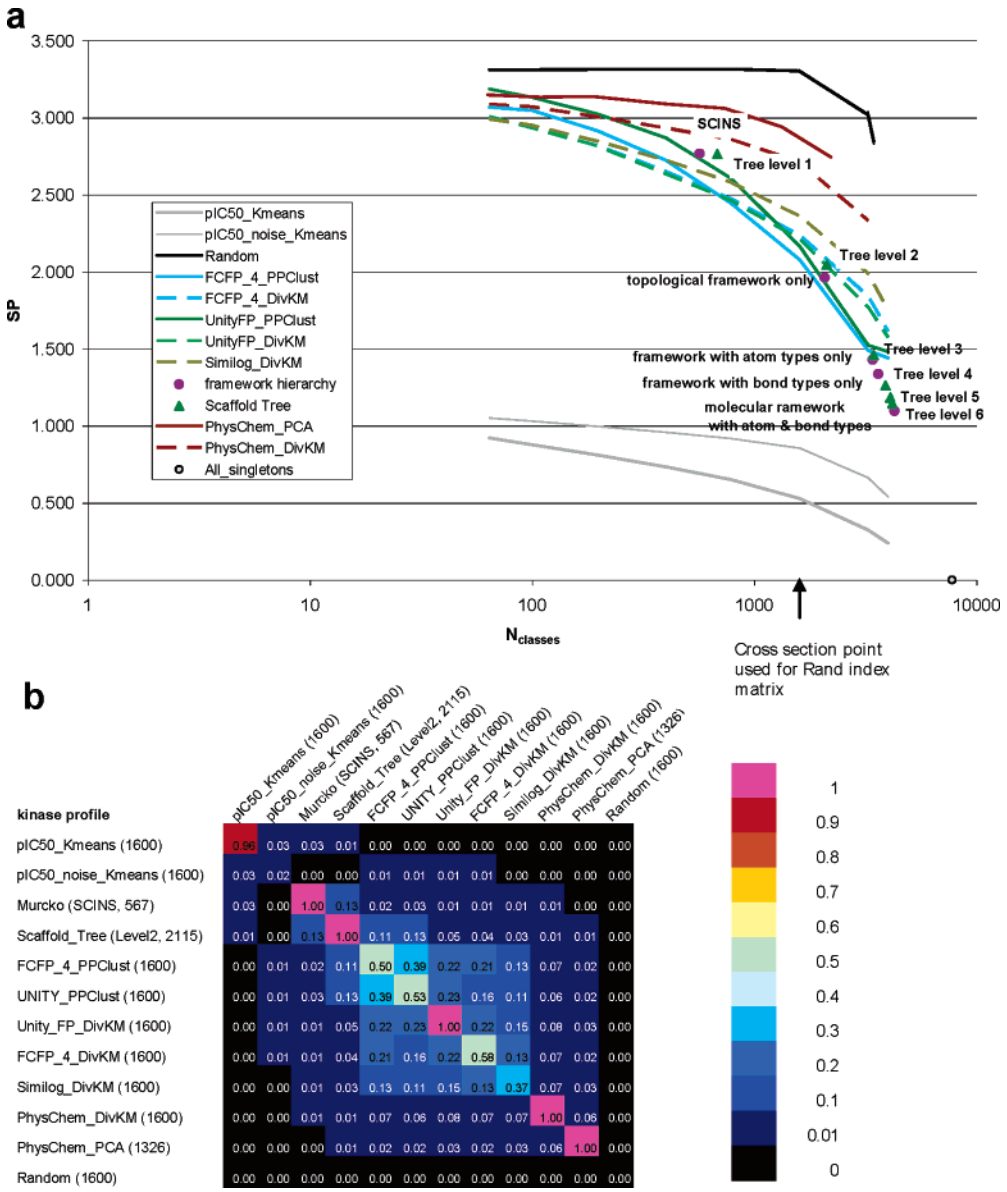

Figure 5. (a) Pareto analysis plot and (b) matrix of adjusted Rand indices for the NCI cancer data subset. For details, see the caption of Figure 3.

Especially interesting are the results obtained with the FEPOPS descriptor. There is almost no difference in each case between the tradeoff curves obtained by the two tie-breaking procedures maj and dist; also, the difference between the tradeoff curves from DivKM and PPClust is only small. Toward higher numbers of clusters, the methods perform worse than all other methods, except the cell-based clustering based on the PCs of the physical chemical properties. With decreasing numbers of clusters, however, the difference in the average spread between the FEPOPS and the fingerprint-based methods diminishes, and in the case of the safety profiling data set, the FEPOPS-based clustering is superior to all other methods when reducing the number of clusters below 20.

Besides common trends in data sets, there are differences, especially between the kinase profile set and the two other data sets. In the kinase profile data set, all scaffold-based methods gave inferior results compared to their performance in the other data sets. On each hierarchy level, there were more molecules per scaffold than in the other data set, indicating that the data set is less diverse in terms of different scaffolds.

Adjusted Rand Indices. As the tradeoff curves of the different clustering methods were rather close together in the Pareto space, it is of interest to investigate whether the methods yield basically the same classifications. This is reflected in the adjusted Rand indices obtained for the comparisons of the classifications resulting from different methods. For each data set, the Rand indices were calculated for a number of clusters, where DivKM and PPClust gave rather similar spreads. For the safety profile, the kinase profile, and the NCI cancer data set, the number of clusters was set to 196, 400, and 1600, respectively. For methods where the number of clusters was not fully controllable, the solution with the number of clusters closest to these values was chosen. Figures 3b, 4b, and 5b show the adjusted Rand indices for each pair of methods in the off-diagonal elements. The values are averaged over the five runs for each method, if the methods are nondeterministic. The diagonal elements show the adjusted Rand index values for different runs of

the same methods, averaged over all pairs of different runs of the respective method.

In all three data sets, the random partitions were, as expected, completely uncorrelated with the biological profile clustering and the structural clusters as well as between the different runs of random partitioning. Also, the biological clustering data show only minimal correlation to the different chemical clustering data. In the case of the safety and kinase profiles, the clusters obtained with the perturbed biological data are still well-correlated with the clusters obtained on the original data. This is not the case with the NCI cancer data subset where the perturbed data show only minimal correlation to the original data, and also the different perturbation runs were only minimally correlated.

In all data sets, there are remarkable differences in the Rand indices between the classifications obtained in different runs of the nondeterministic methods. Whereas the DivKM clustering with the UNITY fingerprints behaves de facto deterministically, and many methods like PPClust with both UNITY and FCFP_4 fingerprints still have rather high adjusted Rand indices, sometimes only a very low correlation between different runs is obtained. This is especially remarkable in the case of PPClust using FEPOPS, although even in this case, the standard deviation for the spread obtained in different runs was not exceptionally high and still on the order of magnitude of 0.01. In the case where FEPOPS descriptors are clustered with DivKM, the correlation between the individual runs is higher, especially when the maj tie-breaking is used. Also, the clustering of the Similog keys with DivKM and the clustering of the RDF codes with SOM have only a comparatively low correlation between the different runs.

When the classification methods are compared, the clustering method seems to have a higher impact than the descriptor. In the case of FCFP_4 and UNITY fingerprints and DivKM and PPClust, the solutions pairs using the same clustering algorithm had a higher adjusted Rand index than the pairs using the same descriptor, but different clustering methods. The two rule-based methods yield rather different classifications having an adjusted Rand index between them of 0.1−0.2 for the three data sets.

DISCUSSION

Scaffold and Scaffold Hopping. Depending on the number of classes which are acceptable, the Pareto frontier of biologically optimal clustering is formed by different classification methods. For high numbers of classes, classification by scaffold tends to be the optimal method, whereas for smaller numbers of classes, 2D fingerprint-based clustering methods are Pareto-optimal. For, even smaller numbers of classes of 3D descriptor-based clustering methods are in some cases displacing the 2D fingerprint-based methods in the Pareto frontier. This is in line with the concept of scaffold hopping, which acknowledges that scaffolds indeed convey information about biological activity, but there are possible forms of generalization beyond the scaffold, which can be captured in suitable descriptors. In this context, it is noteworthy that even in cases where scaffold-based methods performed equally as well as descriptor-based methods—as is the case for those numbers of partitions for which we calculated the Rand index matrix—the scaffold-based classifications are only weakly correlated to the classifications generated by descriptor-based clustering. In this way, different scaffolds can be recognized as similar, and thus scaffold hopping can be performed. However, especially toward the lower number of clusters, the Pareto frontier of chemical clustering is closer to that of the random partitioning than the tradeoff curve obtained by clustering with the biological profiles. This suggests that scaffold hopping is a difficult task, and pairs of biologically similar molecules are often not recognized as similar by chemical structure-based methods.

Considering that "biologically similar" can mean a common absence of any biological activity in all assays included in the profile, the biological clustering as a benchmark has to been seen rather as a theoretical benchmark than one which is practically in reach. The failure to group together inactive molecules in one cluster is most likely also the reason for the small total correlation expressed as the adjusted Rand index between biological clustering and any chemical structure-based partitioning solution. It also provides a good explanation why diversity selections obtained by clustering are not generally beneficial to reduce the number of compounds in HTS and still maintain the number of discovered active classes.[35]

While in the data sets scaffold-based classification is biologically meaningful, one needs to be very careful interpreting this relation. One might assume that the scaffold indeed conveys the activity itself or, at least, provides proper spatial orientation for the side chains in order to display the pharmacophore. However, the relationship between the scaffold and biological activity might also be a consequence of a directed chemical evolution of the scaffold by the synthetic chemist, who will generally, once some initial structure−activity relationship for a scaffold has been established, derive structures for new compounds to synthesize from those compounds which have shown the desired activity. Thus, the chemical space of a scaffold is extended in a directed way toward a certain activity, and other potential substitution patterns of the scaffold are less likely to be explored. This kind of scaffold evolution has likely happened in the case of the kinase set, where in particular, the pharmacophoric features required for interaction with the ATP-binding site common to all kinases are well-documented.[36] This is even more true as at Novartis kinase profiling was mostly only performed once some activity on at least one kinase was discovered or if the compound was synthesized in a medicinal chemistry program directed to a kinase, meaning that this data set consists mostly of kinase-directed molecules. In contrast to this, the targets of the safety-profiling data set are usually perceived as counter-targets, on which one wishes to avoid activity: most of the molecules have not been synthesized with the intention of generating activity on these targets, and therefore, directed chemical evolution to create activity on these targets is less likely. However, even if the relationship between the scaffold and biological activity is a result of directed chemical evolution, the information contained in the scaffold is still of practical value, especially in a setup where one deals with compound sets which are not dominated by nontargeted combinatorial synthesis efforts.

Since all of the data sets used in this study were rather small compared to a corporate screening collection, the

question arises regarding the extent to which the results can be extrapolated to larger compound sets. As despite a 7-fold increase of the data set size between the safety pharmacology profile and the NCI cancer data set similar trends in the Pareto plots have been observed, this indicates that data set size alone allows the extrapolation of the results. However, the distribution of compounds over chemical series and the target panel studied will also influence the outcome of extrapolation to a whole industrial screening collection. In the event where the set contains large chemical series, these could be detected by rule-based classification. It seems to be difficult to predict whether these scaffold-based classifications reflect the activity classes well or if the scaffold-hopping potential of descriptor-based clustering is required to group compounds with similar activity, but different scaffolds, into one class.

Differences Resulting from the Type of Biological Data. The three data sets used in our study covered different types of biological activity, whereas both pharmacology-safety and kinase profiles contained either biochemical or cell-based assays measuring the activity of the compound on a single protein: the NCI cancer data set comprises a panel of phenotypic assays measuring the effect of the compound on cell growth. One might imagine that cell-growth inhibition, which can be effected by different biological pathways, would be much more difficult to relate to chemical structure than the interaction with a single protein target, and therefore, we initially expected that classification based on chemical structures would be less relevant with this type of biological activity. However, this turned out not to be the case, at least not to the extent we had expected.

If there were differences between the data sets, it was rather the kinase profile data set, which differed from the other two sets and yielded with the divisive *K*-means clustering classification results which were halfway between biological and random partitioning even for a smaller number of classes. This suggests that even in binding to the same target binding site in the safety-pharmacology assays there might be at least partially different pharmacophores involved, whereas in the case of the kinases, most ligands are likely to share the common ATP-site binding pharmacophore. GPCRs, which form the majority of assays in the safety-pharmacology profile set, are regarded as rather flexible[37] and can be expected to tolerate more variance in the pharmacophores of their ligands. FEPOPS descriptors which describe the spatial arrangement of the ligand in a rather fuzzy way can be expected to deal better with the topological variance, which is more likely to occur in ligands for more flexible proteins. This might explain why in the case of the safety pharmacology data they are slightly superior to chemical fingerprints when only a small number of partitions is required, whereas the more rigid pharmacophore of the ATP-binding site for kinases can be more easily reduced to substructural elements contained in the fingerprints, which therefore always yield superior classification solutions compared to FEPOPS.

Reproducibility in Clustering. In contrast to the deterministic, rule-based classification methods, clustering methods can produce a different outcome with each run, and with many clustering methods, this is actually the case, as shown by the low Rand indices between the different runs of the same method. This nondeterministic behavior of the cluster can result from ties in the distance metric between two descriptor vectors. Especially in the case of binary fingerprint descriptors, such ties are likely to occur, as MacCuish et al.[38] have demonstrated. In the case of the clustering based on the FEPOPS descriptor, there is another source of nondeterministic behavior, as the FEPOPS descriptor vectors of seven conformer representations per molecule had been clustered, whereas in the case of the other descriptors, each molecule was represented by only one single descriptor vector. Whenever different conformers belong to different clusters, a decision procedure had to be applied that selected one cluster to assign the compound to in order to obtain a nonfuzzy clustering comparable with the other partitioning methods. Here, the tie-breaking procedures we introduced are not yet optimal.

Despite the nondeterministic behavior of the clustering, it needs to be stressed that the different results obtained at each run were found to have equal biological relevance and have almost identical spreads in the Pareto plots. In this context, it is worth noting that the number of different possible classes described by the Stirling number of the second kind[39] is enormous even for data sets of a modest size; for the classification of the safety profile data set with 1006 members into 196 nonempty classes, there are 6.294×10^{1939} solutions. This makes it very likely that there exists a large number of solutions which reflect the true biological clustering of the compounds to a moderate degree as chemical structure-based clusters do. In practical applications, the nondeterministic nature of the clustering algorithms will often not become manifest, as long as the sequence of the records in the input is not changed, because the configuration of many clustering applications forces a constant random seed, unless the user specifies otherwise, and therefore, they seem to behave deterministically to the innocent user. What kind of practical value can clustering beyond the conservative area of chemical scaffolds have, besides being a tool to assess the scaffold-hopping capabilities of a descriptor? For the interpretation of screening data, the usage of clustering as an idea rather than a rule generator can help to spot common activity patterns across chemical classes, and in this aspect, the usage of different clustering procedures may help to detect different aspects of chemical similarity. For the challenge of sampling large data sets with small numbers of compounds, it is clear that structural classification or clustering will not give us "the" optimal solution but will still prevent choosing a sampling solution with clearly redundant compounds.

CONCLUSIONS

We have compared the performance of different classification methods based on chemical structures with respect to the relevance for biological activity. We found that none of the methods analyzed can explain the full spectrum of biological activity contained in the data set. However, both rule-based, scaffold-oriented classification methods and chemical descriptor-based clustering yield results which are at least partially biologically meaningful. A general superiority of one of the methods over the other was not found; however, scaffold-based classifications were usually superior when larger numbers of partitions are acceptable, whereas smaller numbers of classes are best created with clustering using chemical descriptors. The smaller the intended number

of classes, the more the descriptor needs to abstract from substructural features.

ACKNOWLEDGMENT

We acknowledge the work of all Novartis screeners contributing to the profiling data sets used in this study. We thank Stefan Wetzel, Marcus A. Koch, and Herbert Waldmann from the Max Planck Institute of Molecular Physiology in Dortmund, Germany, who co-developed with us the Scaffold Tree algorithm,[31] for many inspiring discussions on the topic of chemical scaffolds.

APPENDIX: SCINS

SCINS describes a reduced graph of the scaffold similar to those used in the MEQI-System described by Xu and Johnson.[30] Each reduced scaffold graph is characterized by a string of numbers in the format **ABCDE-FGHI-JKLM**. These numbers stand for the following:

A. Number of Chain Assemblies. Chain assemblies are contiguous linkers between ring assemblies. They are uncovered by removing all ring bonds in the molecule.

B. Number of Chains. Chains are all unbranched linkers needed to cover all nonring bonds in the molecule.

C. Number of Rings.

D. Number of Ring Assemblies. Ring assemblies are fragments remaining when all acyclic bonds have been removed.

E. Number of Bridge Bonds. A contiguous path of more than one bond shared between more than one rings counts as bridge bond.

F. Number of Ring Assemblies Consisting of Exactly One Ring.

G. Number of Ring Assemblies Consisting of Exactly Two Rings.

H. Number of Ring Assemblies Consisting of >3 Three Rings.

I. Number of Macrocycles.

J. Binned Length of Shortest Chain. If the binned length of the shortest chain exists, it is used; otherwise, it is zero.

K. Binned Length of Second Shortest Chain. If the binned length of the second shortest chain exists, it is used; otherwise, it is zero.

L. Binned Length of Third Shortest Chain. If the binned length of the third shortest chain exists, it is used; otherwise, it is zero.

M. Binned Length of Fourth Shortest Chain. If the binned length of the fourth shortest chain exists, it is used; otherwise, it is zero.

For the binning of chain lengths, the following scheme is applied:

$$\begin{bmatrix} 1 & 2 & 3{,}4 & 5{,}6 & >=7 \\ 1 & 2 & 3 & 4 & 7 \end{bmatrix}$$

Supporting Information Available: Summary statistics on the three data sets is available as Supporting Information. In addition, for the subset extracted from the NCI cancer data set, activity data and clustering results with Scaffold Tree are given. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Dean, P. M. Molecular Recognition: The Measurement and Search for Molecular Similarity in Ligand−Receptor Interaction. In *Concepts and Applications of Molecular Similarity*; Maggiora, G. M., Johnson, M. A., Eds.; John Wiley & Sons: New York, 1990; pp 99−117.

(2) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(3) Krejsa, C. M.; Horvath, D.; Rogalski, S. L.; Penzotti, J. E.; Mao, B.; Barbosa, F.; Migeon, J. C. Predicting ADME Properties and Side Effects: The BioPrint Approach. *Curr. Opin. Drug Discovery Dev.* **2003**, *6*, 470−480.

(4) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biological Spectra Analysis: Linking Biological Activity Profile to Molecular Structure. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 261−266.

(5) Barbosa, F.; Horvath, D. Molecular Similarity and Property Similarity. *Curr. Top. Med. Chem.* **2004**, *4*, 589−600.

(6) Böcker, A.; Schneider, G.; Teckentrup, A. NIPALSTREE: A New Hierarchical Clustering Approach for Large Compound Libraries and Its Application to Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 2220−2229.

(7) Böcker, A.; Derksen, S.; Schmidt, E.; Teckentrup, A.; Schneider, G. A Hierarchical Clustering Approach for Large Compound Libraries. *J. Chem. Inf. Model.* **2005**, *45*, 807−815.

(8) Renner, S.; Schneider, G. Scaffold Hopping Potential of Ligand Based Similarity Concepts. *Chem. Med. Chem.* **2006**, *1*, 181−185.

(9) Brown, N.; Jacoby, E. On Scaffolds and Hopping in Medicinal Chemistry. *Mini Rev. Med. Chem.* **2006**, *6*, 1217−1229.

(10) Downs, G. M.; Barnard, J. M. Clustering Methods and Their Uses in Computational Chemistry. *Rev. Comput. Chem.* **2002**, *18*, 1−40.

(11) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.

(12) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(13) Cases, M.; Garcia-Serna, R.; Hettne, K.; Weeber, M.; van der Lei, J.; Boyer, S.; Mestres, J. Chemical and Biological Profiling of an Annotated Compound Library Directed to the Nuclear Receptor Family. *Curr. Top. Med. Chem.* **2005**, *5*, 763−772.

(14) Cheng, Y. C.; Prusoff, W. H. Relationship between the Inhibition Constant (Ki) and the Concentration of Inhibitor which Causes 50 per cent Inhibition (I50) of an Enzymatic Reaction. *Biochem. Pharmacol.* **1973**, *22*, 3099−3108.

(15) Kelley, L. A.; Gardner, S. P.; Sutcliffe, M. J. An Automated Approach for Clustering an Ensemble of NMR-Derived Protein Structures into Conformationally Related Subfamilies. *Protein Eng.* **1996**, *9*, 1063−1065.

(16) Gillet, V. J.; Willett, P.; Fleming, P. J.; Green, D. V. S. Designing Focused Libraries Using MoSELECT. *J. Mol. Graphics Modell.* **2002**, *20*, 491−498.

(17) Handschuh, S.; Gasteiger, J. The Search for the Spatial and Electronic Requirements of a Drug. *J. Mol. Model.* **2000**, *6*, 358−378.

(18) Hubert, L.; Arabie, P. Comparing Partitions. *J. Classif.* **1985**, *2*, 193−218.

(19) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Topological Descriptors for Similarity-Based Virtual Screening Using Multiple Bioactive Reference Structures. *Org. Biomol. Chem.* **2004**, *2*, 3256−3266.

(20) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391−405.

(21) Jenkins, J. L.; Glick, M.; Davies, J. W. A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes. *J. Med. Chem.* **2004**, *47*, 6144−6159.

(22) Selzer, P.; Ertl, P. Applications of Self-Organizing Neural Networks in Virtual Screening and Diversity Selection. *J. Chem. Inf. Model.* **2006**, *46*, 2319−2323.

(23) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D-Atomic Coordinates for Organic Molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547.

(24) Wild, D. J. New Techniques for Clustering and Analyzing Large Volumes of Chemical Information. *Abstracts of Papers*, 36th Central Regional Meeting of the American Chemical Society, Indianapolis, IN, June 2−4, 2004; American Chemical Society: Washington, DC, 2004.

(25) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinheim, Germany, 1999.

(26) Clark, R. D. OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181−1188.
(27) Sutherland, J. J.; Weaver, D. F. Development of Quantitative Structure−Activity Relationships and Classification Models for Anticonvulsant Activity of Hydantoin Analogues. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1028−1036.
(28) Bayley, M.; Willett, P. Binning Schemes for Partition-Based Compound Selection. *J. Mol. Graphics Modell.* **1999**, *17*, 10−18.
(29) Mason, J. S.; MacLay, I. M.; Lewis, R. A. Application of Computer-Aided Drug Design Techniques to Lead Generation. In *New Perspectives in Drug Design*; Dean, D. M., Jolles, G., Newton, C. G., Eds.; Academic Press: London, U. K., 1995; pp 225−253.
(30) Xu, Y. J.; Johnson, M. Using Molecular Equivalence Numbers To Visually Explore Structural Features that Distinguish Chemical Libraries. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 912−926.
(31) Schuffenhauer, A.; Ertl, P; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree − Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47*, 47−58.
(32) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting Biologically Relevant Chemical Space: A Structural Classification of Natural Products (SCONP). *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17272−17277.
(33) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem. A* **1998**, *102*, 3762−3772.
(34) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.
(35) Schuffenhauer, A.; Brown, N.; Selzer, P.; Ertl, P.; Jacoby, E. Relationships between Molecular Complexity, Biological Activity, and Structural Diversity. *J. Chem. Inf. Model.* **2006**, *46*, 525−535.
(36) Traxler, P.; Furet, P. Strategies toward the Design of Novel and Selective Protein Tyrosine Kinase Inhibitors. *Pharmacol. Ther.* **1999**, *82*, 195−206.
(37) Jaakola, V.-P.; Prilusky, J.; Sussman, J. L.; Goldmann, A. G Protein-Coupled Receptors Show Unusual Patterns of Intrinsic Folding. *Protein Eng., Des. Sel.* **2005**, *18*, 103−110.
(38) MacCuish, J.; Nicolaou, C.; MacCuish, N. E. Ties in Proximity and Clustering Compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 134−146.
(39) Goldberg, K.; Newman, M.; Haynsworth, E. Combinatorial Analysis. In *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, 10th ed.; Abramowitz, M., Stegun, I. A., Eds.; U.S. Government Printing Office: Washington, DC, 1972; pp 824−825.

CI6004004