CerBeruS: A System Supporting the Sequential Screening Process†

Michael F. M. Engels,*,‡ Theo Thielemans,§ Danny Verbinnen,⊥ Jan P. Tollenaere,*,‡ and Rudi Verbeeck⊥

Departments of Theoretical Medicinal Chemistry, Research Support, and Global Information Management, Janssen Research Foundation, Turnhoutseweg 30, B-2340 Beerse, Belgium

Received June 29, 1999

This paper describes the general design and application of CerBeruS, a computer-based system for supporting the process of sequential screening. CerBeruS stands for cluster-based selection, with cluster analysis forming the pivotal part of the system. CerBeruS uses the Ward's clustering method for partitioning the data set to be screened into smaller, more homogeneous subsets. One representative is picked from each subset and suggested as a screening candidate. Although the number of compounds submitted to screening is most often driven by the capacity of the assay, CerBeruS provides a statistical measure that computes the optimal number of clusters in the data set. This measure forms a point of reference for all screening experiments. Different hierarchies of subsets are stored in an Oracle database. Information about the size and content of a cluster can be retrieved from this database via a Visual Basic application. How these components work together in the CerBeruS system is demonstrated on a large data set. In addition, we show that, using the statistical measure, one can find an optimal trade-off between screening effort and number of hits.

INTRODUCTION

Screening of large compound libraries has been established as a key component of the pharmaceutical lead-identification process in many pharmaceutical companies. The goal is to identify compounds in the company or external compound collections that show activity against a particular biological target. Compounds that show appropriate activity may ultimately form the basis of a lead optimization program aimed at optimizing not only the biological activity but also pharmacokinetic, physicochemical, and pharmaceutical properties by modification of the chemical structure.

The sequential screening experiment, represented in Figure 1, is one of the most popular types of screening experiments. It is applied in situations in which screening the complete set of compounds is not a feasible option, for example because the capacity of the assay is limited or such screening is not cost-effective.

The initial screening experiment starts with the definition of an initial subset of compounds taken from the compound library under investigation. This set of compounds, referred to as initial sample,[3] is submitted to a first screening run. Provided that hits are picked up by this first screening run, a data analysis step follows aiming at the identification of structural features relevant to the biological activity. On the basis of this first structure−activity relationship (SAR), a database search is performed with a view to identifying additional compounds that have not been tested yet. These compounds form an additional sample, which is subject to a further screening run, etc. This circle of testing−analyzing−testing is pursued until a strong structure−activity relationship has been established.

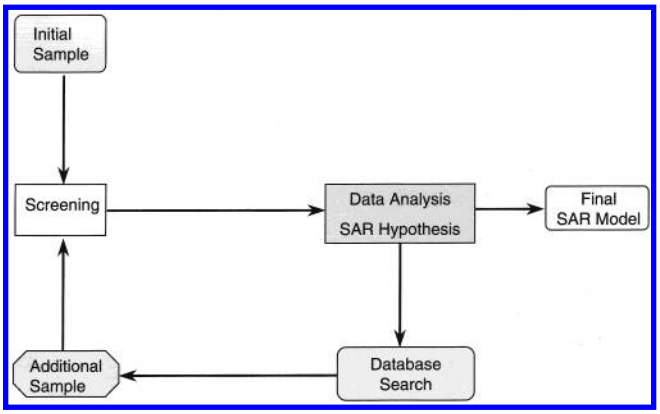

Figure 1. Flow chart of the sequential screening experiment.

Given the ease today with which potential targets are cloned, expressed, and purified, it is needed not only to optimize the throughput of screening systems but also to improve the efficiency of the screening experiment per se. In this paper we introduce CerBeruS, an integrated system which supports the sequential screening process. We have developed this system at the Janssen Research Foundation in an effort to optimize the sequential screening experiment in its entirety and to improve the workflow between some of its components. CerBeruS supports three components of the sequential screening experiment: it assists in defining appropriate initial samples for screening; it rapidly identifies those compounds in the data set which have a high probability of being active and which have not been tested yet; it supports the process of establishing SAR(s).

We shall now describe the general design of the CerBeruS system and demonstrate how it is used on a large data set for which screening biological data are available.

* To whom corrrespondence should be addressed.
† Presented at the Fifth International Conference on Chemical Structures, June 6-19, 1999, Noordwijkerhout, The Netherlands.
‡ Department of Theoretical Medicinal Chemistry.
§ Department of Research Support.
⊥ Department of Global Information Management.

10.1021/ci990435+ CCC: $19.00
Published on Web 12/21/1999

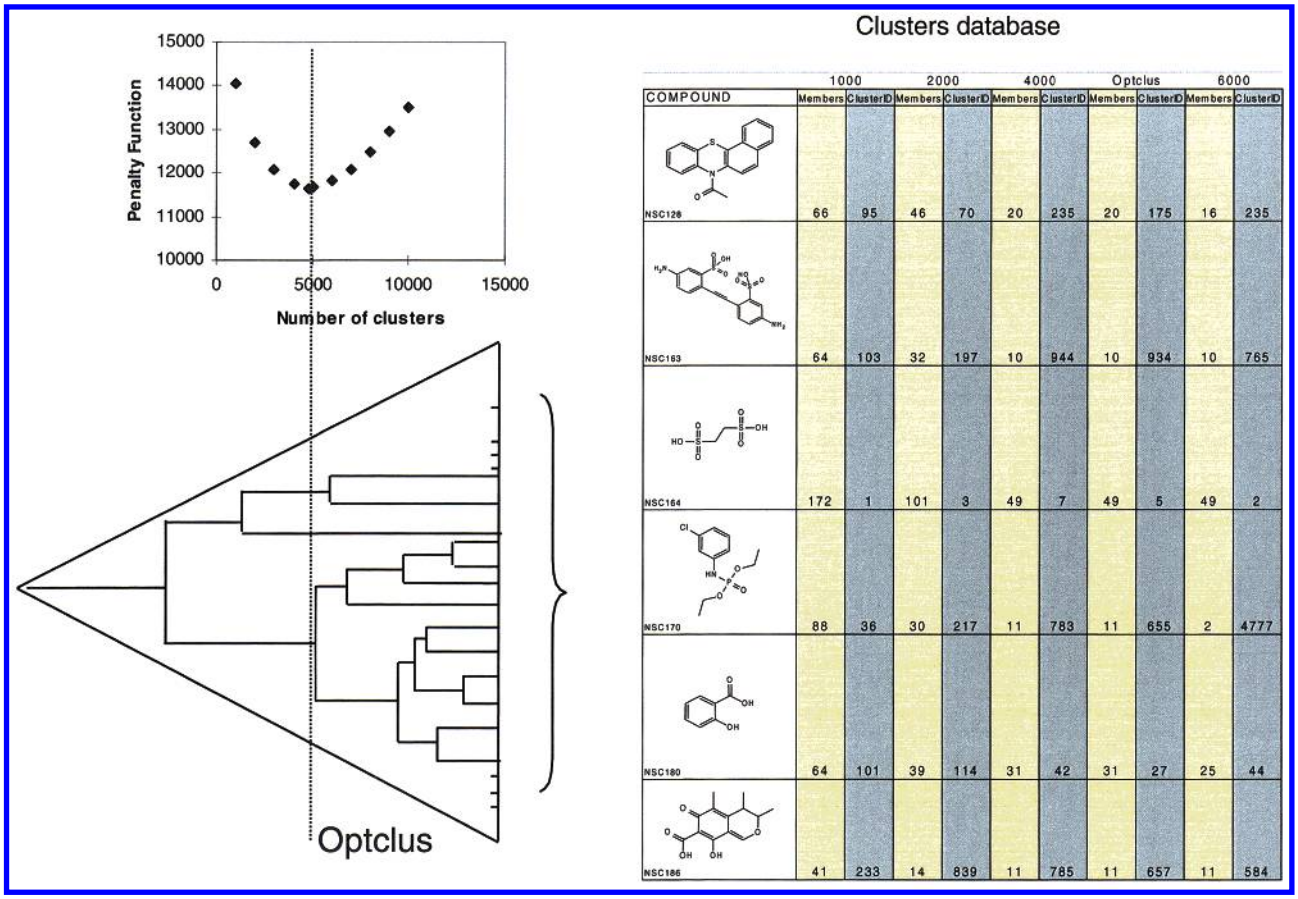

Figure 2. Example of the information stored in the clusters database.

CERBERUS

Rationale. Hierarchical cluster analysis is one of the most popular methods for defining initial samples for screening.[1,2] In particular, Ward's clustering[3] in combination with two-dimensional topological descriptors or fingerprints has proven to outperform alternative methods for selecting representative subsets from a compound library.[4] Following Bayada et al.,[4] the data set is partitioned into smaller, more homogeneous subsets and one representative is picked from each subset on the basis of a reproducible criterion, for example compounds closest to the cluster centroid. This procedure has now been established at several pharmaceutical companies.[1,2,4,5]

There are two further advantages of using hierarchical clustering in the context of sequential screening. First, an additional sample for subsequent screening runs can be easily extracted from a cluster solution when an active compound is found.[2] In that case, the whole cluster from which the representative has been picked is submitted to a second screening run. Because all members of a given cluster are structurally related to their cluster representative, they possess a greater likelihood of being active. Second, given the results of the first and all subsequent screening runs, rapid SAR analyses can be carried out by studying the cluster dendrogram at different levels of the cluster hierarchy. In particular, clustering can suggest not only the grouping of the hits into distinct chemical families, but can also assist in identifying structural properties relevant for the biological activity concerned.

General Design of CerBeruS. CerBeruS was originally conceived with the intention to speed up the data analysis step within the sequential screening procedure by combining several of the above-described aspects into a single integrated system. CerBeruS uses Ward's clustering for establishing a "pedigree" for all compounds of the data set under investigation. Several cut points are identified in the cluster hierarchy around a certain point of reference (see below), and the corresponding clusters are stored in an Oracle database (the clusters database) for fast retrieval (Figure 2). The compound identification (ID), the size of the respective cluster, the distance of the compound from the centroid of the cluster, and the compound ID of the cluster representative are stored in Oracle tables, with the cluster hierarchy and the cluster identification number being used as keys.

In order to optimally design our database for screening, we use the penalty function introduced by Kelley et al.[6] for determining the most suitable cut point through the cluster hierarchy. Kelley's penalty function, P_i, is a trade-off between the number of clusters, $Nclus_i$, and the normalized average spread over all clusters, $ASpread_i^{norm}$, at stage i of the cluster hierarchy. P_i is defined as

$$P_i = ASpread_i^{norm} + Nclus_i$$

The normalized average spread at stage i is defined as

$$ASpread_i^{norm} = \left(\frac{N-2}{\max(ASpread) - \min(ASpread)}\right) \times (ASpread_i - \min(ASpread)) + 1$$

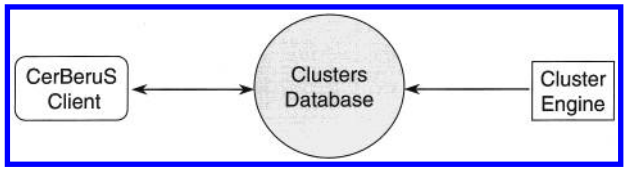

Figure 3. Components of the CerBeruS system.

where $ASpread_i$ is the average over the spread, i.e. the mean intercompound distance within each cluster, of all clusters at cluster hierarchy *i*, and min(*ASpread*) and max(*ASpread*) are the minimum and maximum values of the *ASpread* values across all clusters of this cluster level. Singletons are not taken into account when *ASpread* values are calculated. *N* corresponds to the total number of structures in the data set. This normalization procedure ensures that the set of $ASpread_i$ values falls within the range 1 to $N - 1$.

The penalty function is calculated for all levels of the cluster hierarchy. At the level with the smallest penalty value, there is a trade-off between the normalized average spread and the number of clusters, which corresponds to maximum homogeneity within the clusters, while the number of clusters is kept limited. CerBeruS uses the cluster solution suggested by Kelley's penalty function as a point of reference. Around this reference point other cut points are identified as described above.

The general design of CerBeruS implies that the database that is used for screening can be partitioned into clearly distinct clusters of chemical compounds. This assumption seems to be justified for compound libraries, which grew from an effort to optimize the SARs of several lead compounds.

Components of the CerBeruS System. CerBeruS consists of three main components (see Figure 3): the Oracle database, which stores cluster solutions at selected levels of the cluster hierarchy; the CerBeruS client, a Visual Basic application that forms the interface with the database; the clustering engine, which performs Ward's clustering.

Technical Implementation. CerBeruS uses Daylight CIS[7] fingerprints as molecular descriptors. The clustering is performed with the BCI[8] implementation of Ward's algorithm. This implementation uses the squared Euclidean distance as intercompound distance measure. The Kelley penalty function is calculated with the OPTCLUS program of BCI.[9]

APPLICATION EXAMPLES

Data Set. The functionality and capability of the CerBeruS system are demonstrated with a publicly available data set from the National Cancer Institute.[10] The original data set consists of 32 110 compounds (version May 1997), which were tested for evidence of anti-HIV activity. The anti-HIV activity of these compounds was measured by means of a cell-based assay,[11] and the activity of a compound was assigned to one of three categories: confirmed active (CA), confirmed moderately active (CM), and confirmed inactive (CI). We did not use the three classes; for us, compounds were either active (=CA and CM) or inactive (CI). Since we excluded heavy metal compounds from the analysis, the final data set consists of 30 920 compounds, 650 of which were active. The data set was clustered and the optimal cluster level determined. The optimal number of clusters was 4768. Around the Optclus level—the point of reference—further cuts were performed in the cluster hierarchy at the following levels: 1000, 2000, 4000, 6000, and 8000 clusters. The different cluster solutions were downloaded into the Oracle database.

Table 1. Hit Rates of the First Run of the Virtual Sequential Screening Experiment Obtained at Different Levels of the Cluster Hierarchy

	level					
	1000	2000	4000	Optclus	6000	8000
hits (1st run)	23	47	90	97	117	156
hit rate, %	2.3	2.4	2.3	2.0	2.0	2.0

An Example Session. Here, we demonstrate the use of the CerBeruS client (Figure 4). Of the 97 active compounds at the Optclus level, 6 form the input for a CerBeruS session: NSC633810 (TIBO), NSC667488, NSC646436, NSC657150, NSC624486, and NSC65852. All six compounds are representatives of their clusters; i.e., they are closest to the centroid of the cluster.

Queries are entered in the "Input and Filter Workbench". Filters can be set on availability of the compound (stock) or solubility.

The "Analysis Workbench" shows the number of compounds in each of the requested clusters at several levels of the cluster hierarchy. The CerBeruS client allows sorting the columns on these numbers consecutively from the left to the right side of the analysis matrix. This sorting procedure reestablishes the underlying cluster hierarchy, which can be used for grouping compounds into different families or for SAR analysis. For example, according to the analysis workbench the compounds NSC667488, NSC646436, and NSC657150 belong to one cluster at the hierarchy level 1000. This cluster consists in total of 32 compounds. At the next higher level of the cluster hierarchy (level 2000), the three compounds are distributed over two clusters. NSC667488 and NSC646436 still belong to the same cluster; NSC657150 is split off from that group and assigned to another cluster.

Testing Kelley's Penalty Function. The initial idea for implementing a point of reference was that we could increase the efficiency of the sequential screening experiment as a whole. To test this idea, we performed several virtual screening experiments at different levels of the cluster hierarchy. The virtual screening experiment consists of a first and a second screening run. In the first screening run, a representative for each cluster is chosen and "tested". As representative we picked the compounds closest to the centroid of the cluster. Table 1 shows the results of these screening runs as a function of the cluster hierarchy. Active compounds detected in the first screening run define active cluster subsets.[11] All members of the active cluster subsets are submitted to the second virtual screening experiment. The results of this second screening run are shown in Table 2.

To measure the efficiency of the screening experiment we introduce the target function *T* which is defined as

$$T = N_{\mathrm{Act}} - N_{\mathrm{Scr}}$$

T maximizes the number of hits N_{Act} while minimizing the screening effort (number of compounds screened N_{Scr}).

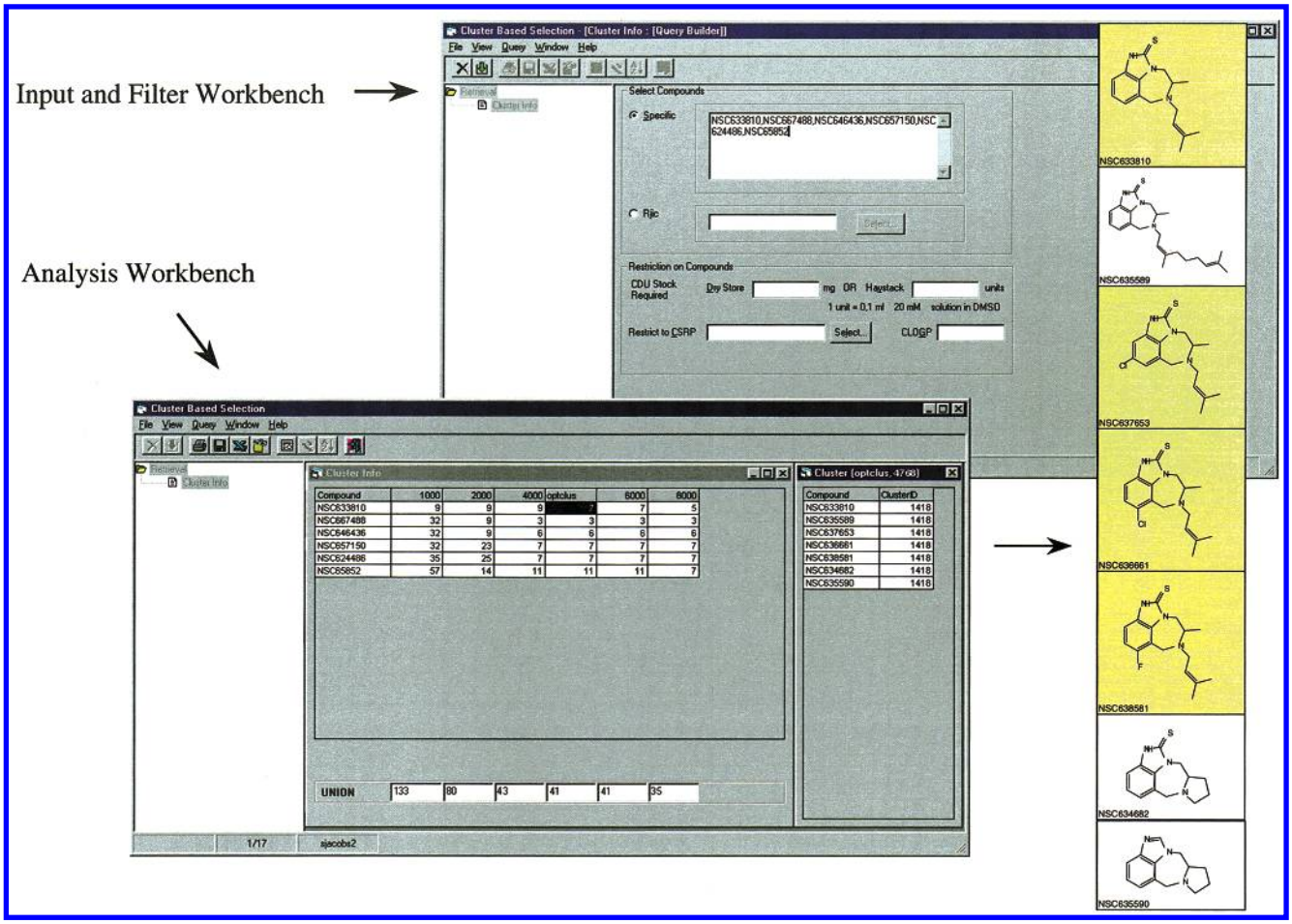

Figure 4. Example session showing the different workbenches of the CerBeruS client application. Compounds are selected by their compound ID and specified in the Input and Filter Workbench. This set of compound IDs forms a query for a database search in the cluster database. The Analysis Workbench yields the result of such a database search, in which for each compound the number of cluster members is shown at different levels of the cluster hierarchy. More explicit information about the members of a cluster can be obtained by double clicking on the data elements of the analysis matrix. In this example session, individual members of the cluster represented by TIBO (NSC633810) at the Optclus (=point of reference) level are retrieved (list on the right side of the Analysis Workbench) and displayed. Compounds that were active in the assay are highlighted in yellow.

Table 2. Hit Rates of the Second Run of the Virtual Sequential Screening Experiment Obtained at Different Levels of the Cluster Hierarchy

	level					
	1000	2000	4000	Optclus	6000	8000
no. compds tested	612	692	616	602	523	504
hits (2nd run)	159	185	187	199	196	194
hit rate, %	25.0	26.7	30.4	33.1	30.6	38.5

So that the two quantities can be compared in relative units, they are normalized by their respective maximum (total number of active compounds [T_{Act}] and total number of compounds in the data set [N_{Data}]).

$$T = \frac{N_{Act}}{T_{Act}} - \frac{N_{Scr}}{N_{Data}} = PA - CR$$

where *PA* and *CR* are the fractions of the active compounds and the coverage rate of the screening experiment, respectively. *T* values for the different hierarchy levels are shown in Table 3.

Table 3 shows that the number of hits increases as more compounds are screened. The efficiency, however, represented by the target function *T*, is at a maximum at the Optclus level, i.e., the point of reference. Note that in practice *T* cannot be calculated, as the total number of active compounds T_{Act} is unknown. This experiment suggests that Optclus is a good approximation for the maximum *T* and, therefore, to the most efficient screening strategy. This conclusion is supported by a recent study conducted by Wild and Blankley.[13] In this study, they tested different clustering hierarchy level selection methods on seven different data sets. None of the tested selection methods performed consistently better across the different data sets. However, based on mean and worst case performance, they suggest that Kelley's penalty function selection method is the most appropiate when used together with Daylight fingerprints.

Table 3. Results for the Complete Screening Experiment

	level					
	1000	2000	4000	Optclus	6000	8000
no. compds tested	1612	2692	4616	5370	6523	8504
no. hits detected	182	232	277	296	313	350
T (target function)	0.228	0.270	0.277	0.281	0.271	0.265

FINAL REMARKS AND CONCLUSION

The number of biologically relevant targets that are submitted to screening experiments is increasing every year.

This development presents a severe challenge to all people involved in the analysis of screening data. CerBeruS has been developed with the intention to streamline some of the working procedures involved in the screening process. Since its introduction into the company, it has been found very useful in particular thanks to its easy-to-comprehend concept, its design, and its functionality. It has been successfully applied as an alternative to existing systems for high-throughput SAR analysis (grouping hits) and database searching (searching for additional samples).

A feature that has proved extremely convenient in working with CerBeruS is the point of reference. The information about the number of chemical series in a database can be related to the minimum number of compounds that the initial sample should include. However, the application of such a point of reference implies that the compounds in the database under investigation are clearly grouped in distinctive clusters. In that respect, we can refer to our first and preliminary experiences. For diverse compound libraries, the definition of a point of reference is useless. In that case, other procedures for subset selection are better suited.[14−17]

REFERENCES AND NOTES

(1) Hodes, L. Clustering a large number of compounds. 1. Establishing the method on an initial sample. *J. Chem. Inf. Comput. Sci*. **1989**, *29*, 66−71.
(2) Dunbar, J. B. Cluster-based Selection. *Perspectives Drug Discovery Des.* **1997**, *7/8*, 51−63.
(3) Ward, J. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, 58, 326−244.
(4) Bayada, D. M.; Hamersma, H.; van Geerestein, V. J. Molecular Diversity and Representativity in Chemical Databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1−10.
(5) Stanton, D. T.; Morries, T. W.; Roychoudhury, S.; and Parker, C. N. Application of Nearest-Neighbor and Cluster Analyses in Pharmaeutical Lead Discovery. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 21−27.
(6) 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.
(7) *Daylight fingerprint manual*, version 4.51; Daylight Chemical Information Systems Inc.: 18500 Von Karman, #450, Irvine, CA.
(8) *BCI Clustering package*, versions 2.5 & 3.0; Barnard Chemical Information Ltd.: 46 Uppergate Road, Sheffield, S6 6BX U.K.
(9) *BCI Optclus program*, version 1.0; *Ibid*.
(10) The AIDS antiviral screen data set is available from the Developmental Therapeutics Program at the National Cancer Institute, http://dtp.nci.nih.gov/
(11) Weislow, O. S.; Kiser, R.; Fine, D. L.; Bader, J.; Shoemaker, R. H. Boyd, M. R. New Soluble-Formazan Assay for HIV-1 Cytopathic Effects: Application to High-Flux Screening of Synthetic and Natural Products for AIDS-Antiviral Activity. *J. Natl. Cancer Inst.* **1991**, *81*, 577−586.
(12) Brown, R. D.; Martin, Y. 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.
(13) Wild, D. J.; Blankley, C. J. A Comparison of 2D Fingerprint Types and Hierarchy Level Selection Methods for Structural Grouping Using Ward's Clustering. Poster Contribution at the Fifth International Conference on Chemical Structures, June 6−10, 1999, Noordwijkerhout, The Netherlands.
(14) Pötter, T.; Matter, H. Random or Rational Design? Evaluation of Diverse Compound Subsets from Chemical Structure Databases. *J. Med. Chem.* **1998**, *41*, 478−488.
(15) Young, S. S.; Farmen, M.; Rusinko, A. Random versus rational which is better for general compound screening? *Network Sci.* [electronic publication] **1996**, *2* (7), http://www.netsci.org/Science/Screening/feature09.html.
(16) Lajiness, M. S. Dissimilarity-based Compound Selection Techniques. *Perspectives Drug Discovery Des.* **1997**, *7/8*, 65−74.
(17) Mason, J. S.; Pickett, S. D. Partition-based Selection. *Perspectives Drug Discovery Des.* **1997**, *7/8*, 85−114.

ACKNOWLEDGMENT

We are very grateful to Steven Van Leemput for his initial contributions. We thank Mathy Froeyen, Sven Jacobs, and Dirk Schellinck for programming the CerBeruS client, and Geoff Downs from Barnard Chemical Information Ltd. for valuable discussions.

CI990435+