Comparison of the NCI Open Database with Seven Large Chemical Structural Databases

Johannes H. Voigt,[†] Bruno Bienfait,[†,§] Shaomeng Wang,[‡] and Marc C. Nicklaus*,[†]

Laboratory of Medicinal Chemistry, Center for Cancer Research, National Cancer Institute, National Institutes of Health, NCI at Frederick, 376 Boyles Street, Frederick, Maryland 21702, and Structural Biology and Cancer Drug Discovery Program, Lombardi Cancer Center & Departments of Oncology and Neuroscience, Georgetown University Medical Center, 3970 Reservoir Road, Washington, DC 20007

Received November 8, 2000

Eight large chemical databases have been analyzed and compared to each other. Central to this comparison is the open National Cancer Institute (NCI) database, consisting of approximately 250 000 structures. The other databases analyzed are the Available Chemicals Directory ("ACD," from MDL, release 1.99, 3D-version); the ChemACX ("ACX," from CamSoft, Version 4.5); the Maybridge Catalog and the Asinex database (both as distributed by CamSoft as part of ChemInfo 4.5); the Sigma-Aldrich Catalog (CD-ROM, 1999 Version); the World Drug Index ("WDI," Derwent, version 1999.03); and the organic part of the Cambridge Crystallographic Database ("CSD," from Cambridge Crystallographic Data Center, 1999 Version 5.18). The database properties analyzed are internal duplication rates; compounds unique to each database; cumulative occurrence of compounds in an increasing number of databases; overlap of identical compounds between two databases; similarity overlap; diversity; and others. The crystallographic database CSD and the WDI show somewhat less overlap with the other databases than those with each other. In particular the collections of commercial compounds and compilations of vendor catalogs have a substantial degree of overlap among each other. Still, no database is completely a subset of any other, and each appears to have its own niche and thus "raison d'être". The NCI database has by far the highest number of compounds that are unique to it. Approximately 200 000 of the NCI structures were not found in any of the other analyzed databases.

INTRODUCTION

Since 1955, the National Cancer Institute (NCI) has collected samples of compounds from both organic synthesis and natural source extracts for testing in anticancer, and more recently anti-AIDS, assays of various types. Concomitant with this effort of collecting and testing samples, computerized information necessary to keep track of all the data associated with this large-scale effort was created starting from a very early stage of the project. One of the datasets that have thus evolved over time is the collection of chemical structures, in some machine-readable form, of the samples from organic synthesis submitted to NCI for testing. This large chemical structure database is often referred to as "the NCI database". Its total size as of the time of this writing (late 2000) is about 500 000 molecular entries. Further history of this database as well as a description of previous projects making use of it can be found in several prior publications.[1−7]

While academic, government, and other nonprofit laboratories have usually provided samples free of any restrictions, industrial laboratories often request and are granted in so-called "Discreet Agreements," confidentiality regarding both the structures themselves and the test results. As a result of these Agreements, access to these "discreet" compounds and the data associated with them is highly restricted and not usually possible for anyone outside NCI's Developmental Therapeutics Programs[8] (DTP), the part of NCI that is responsible for the continuing screening efforts and the NCI database as a whole.

While about half of the NCI database is in this way inaccessible and of little interest to the public, the other half of the structural data is free of any disclosure and usage restrictions and therefore termed "open". These data are completely in the public domain and often referred to as the "Open NCI Database" (or even, somewhat inaccurately, simply as "the NCI database" as far any public application is concerned). DTP has made most of these open data[9] freely available on their Web site.[10] Various companies are offering this database, or parts thereof, in the original or in a processed format, often in conjunction with their chemical database programs. A fully searchable version of the Open NCI Database, enhanced with additional data, that is freely accessible via a Web-based interface has recently been implemented by members of this author group.[11]

NCI maintains a large repository of actual samples of many of the compounds in the structural database. While some samples are obviously one-time supplies from, say, an academic group working on a specific project, and may therefore have been depleted over time, other chemicals are replenished in the repository on an occasional or regular basis. It is therefore possible at any time to find actual

* Corresponding author e-mail: mn1@helix.nih.gov.
[†] Laboratory of Medicinal Chemistry.
[‡] Georgetown University.
[§] Present address: ChemCodes Inc., 1300 Englert Drive, Suite G, Durham, NC 27713.

10.1021/ci000150t CCC: $20.00
Published on Web 03/29/2001

samples available for about 60% of the compounds in the NCI database, with about an equal availability rate for both the discreet and the open part of the database.

While the pace of acquiring new compounds for testing by DTP has slowed in the recent past, the Open NCI Database is still a very useful resource for researchers both within and outside NCI. This is particularly the case since DTP is currently making available a very substantial subset of the open compounds (ca. 140 000, on 96-well microtiter plates) to the scientific public for essentially the shipping costs.[12] Even when samples are not available or of interest, the Open NCI Database constitutes the largest freely available, public domain chemical structural database with its approximately quarter-million entries and is therefore often used in studies where a large chemical database is needed.

However, the Open NCI Database is certainly not the only large chemical database available. The ultimate reference collection of molecular structures is, with little doubt, the Chemical Abstracts Services (CAS) Registry database, attempting to comprise all chemical substances published, irrespective of any unifying theme or availability. However, the CAS Registry as well as CAS's other databases[13] are not usually available in bulk form and are thus of no use where such datasets are needed. Similar limitations, be it technically or by license restrictions, of not being able to obtain, or extract, bulk collections of connection tables apply, to the best of our knowledge, to other very large chemical databases such as SPRESI[14] and Beilstein.[15] Furthermore, their very nature of being very general reference databases may make them less suited for targeted tasks such as pharmacophore searches and other virtual screening approaches in the context of computer-aided drug development. These databases therefore could not be, and thus were not, included in the current comparison.

Smaller, but still large, databases are available of more well-defined origin and/or for more dedicated purposes. These comprise collections of commercially available compounds, be it from compilers of many manufacturer catalogs, or from some of the largest manufacturers themselves; collections of structures and other data of actual drugs used worldwide; and databases of experimentally determined molecular atom coordinates such as crystal or NMR structures.

In the context of drug development, availability of actual samples of compounds for testing is obviously of significant importance. The large databases of commercially available compounds therefore figure prominently among the databases analyzed in this comparison. These databases come in various sizes (ca. 50 000 to >200 000 compounds) and may essentially be the catalog of a single supplier or combinations of such catalogs from tens if not hundreds of chemical manufacturers, compiled by chemical information software and database vendors. Their prices vary widely, ranging from nominal fees for shipping to well over U.S. $10 000 on the current U.S. American market.

A question that has arisen repeatedly in our ongoing drug development research as well as being posed to us by users of the NCI database is that of the overlap of the Open NCI Database with other such large chemical databases. Related to this question are others, such as what the internal duplication rate is for each database, what the diversity is, how similar the databases are to each other ("similarity overlap"), how stereochemistry information, or the lack thereof, may influence the preceding results, and other points of interest.

While a handful of comparisons[16−21] of large databases have been published during the past 5 years or so, with the emphasis mostly placed on the analysis of molecular diversity, only one of these papers appears to have included the NCI database so far. Even this publication used only a somewhat older subset of approximately 127 000 open NCI compounds.[22] Thus, to our knowledge, there exists no published comparison to date of the entire publicly available Open NCI Database with any other large database.

DATASETS AND METHODS

General Approach. To make possible the required structure-by-structure comparisons across all databases, each structural entry in each database was converted into a unique hash code (64-bit encoded) using the CACTVS chemical information toolkit.[23] Using these hash codes, we first established the internal duplication rate for each database. Then, the mutual overlap of the nonduplicate subsets of each database with each other was determined. Because of the lack of any stereochemistry information in the NCI database, a special analysis of the stereochemistry aspect was conducted. In fact, all duplication and overlap analyses were performed in a dual manner by projecting all databases' compounds both onto stereospecific hash codes and nonstereospecific hash codes and using both sets for separate rounds of comparisons. Finally, the diversity of each of the databases was probed, using two different clustering techniques applied to several different descriptors derived for all compounds, and a similarity overlap of all databases with each other was determined based on these descriptors.

Databases. After obtaining the bulk database files from their various sources, each one of these datasets was extracted into, or converted to, a file in SDF (structure-data file) format[24] for the purpose of this study. All database entries that did not include a chemical structure were removed. Table 1 gives an overview over the databases compared in this paper.

The Open NCI Database (labeled "NCI" in the tables) as used in this study was a dataset put together from the publicly available data provided by DTP[10] and was current up to all posted structures up to about October 1999. This dataset comprised about 249 000 structures. (Various versions of it can be downloaded from our Web server.)[25] As mentioned before, no information about absolute stereochemistry and double bond geometry (E−Z) of any compound is contained in the NCI database connection tables.

The Available Chemicals Directory ("ACD," from MDL,[26] release 1.99, 3D-version) is a combination of vendor catalogs and comprised approximately 222 000 compounds. It was used as obtained in SDF format. A problem regarding stereochemistry in this database is that, for racemic compounds and compounds with unspecified stereo centers, one stereoisomer was apparently chosen by random during the 2D to 3D conversion process used to generate the 3D structures, while the "stereocenter unspecified" flag was not set in the resulting SD file with the 3D-coordinates, which would be the usual mechanism to ensure clarity.

Table 1. Databases Compared in This Study ($: 0−1000$, $$: ≥10 000$; Percentages Indicate the Loss of Entries Due to Internal Duplication)

database	price	no. of entries	no. of structures	unique (nonstereospec. hash code)	unique (stereospec. hash code)	E/Z specified	R/S specified
NCI	0	249 081	249 081	235 461 (5.5%)	235 461 (5.5%)	-	-
ACX	$	137 003	113 554	98 318 (13.45)	98 788 (13.0%)	+	+
ASINEX	$	139 861	139 861	137 810 (1.5%)	137 835 (1.4%)	+	-
SIGALCAT	$	132 528	99 491	91 072 (8.5%)	93 243 (6.3%)	+	+
MAY	$	55 281	55 281	55 064 (0.4%)	55 262 (0.02%)	+	-
CSD	$$	(80 442)[a]	80 442	72 680 (9.4%)	74 510 (7.1%)	+	+
WDI	$$	62 754	57 617	53 895 (6.5%)	56 096 (2.6%)	+	+
ACD	$$	221 668	221 668	211 274 (4.7%)	217 216 (2.0%)	+	+[b]

[a] Subset of the complete CSD. Only those compounds selected that had both flag 57 ("organic") and flag 153 ("coordinates available") set. [b] For racemic compounds and compounds with unspecified stereocenters, "stereocenter unspecified" flag not set (see text).

The ChemACX ("ACX," from CamSoft,[27] Version 4.5) database is a compilation of commercially available compounds covering products of over 79 vendors. The ACX version used contained approximately 137 000 entries. The Maybridge Catalog[28] ("MAY") and the Asinex[29] database ("ASINEX," both were part of ChemInfo 4.5 from CamSoft, included as separate database files) contain commercially available libraries, with the Asinex mostly of combinatorial chemistry origin. The former database contained about 55 000 structures in the version used, the latter approximately 140 000 compounds. ACX, MAY, and ASINEX were exported from the ChemFinder program (v. 4.5). For MAY and ASINEX the stereocenters are unspecified, and the respective flag is set accordingly in the SD file.

The Sigma-Aldrich[30] Catalog ("SIGALCAT," CD-ROM, 1999 Version) was exported from the search software provided on the CD-ROM obtained from the company. All structures lack charge assignments, and some structures contain wrongly assigned covalent bonds between ionic fragments. This causes calculation of the wrong hash codes for ionic compounds.

The World Drug Index[31] ("WDI," from Derwent, version 1999.03), a database of about 63 000 drugs and pharmacologically active compounds, including all drugs marketed worldwide, was included for reason of high interest in the context of drug development. It was acquired in SDF format.

Finally, the organic part of the Cambridge Crystallographic Database ("CSD," from the Cambridge Crystallographic Data Center[32] (CCDC), 1999 Version 5.18) was included in our comparison as the largest database of experimental 3D-structures of small molecules. All organic molecules which had coordinates (flags 57 and 153) in the CSD were exported in Sybyl mol2 format from CCDC's program Quest and converted into SDF format. This yielded about 80 000 structures. All cocrystallized solvent molecules and multiple copies of the same molecule in the asymmetric unit cell were removed using a CACTVS script.

Hash Codes. A hash code is usually the highly compressed encoding of a data structure—in this case the molecular structure—to one value within a fixed range, which can therefore be represented on a computer with a fixed bit length. If a hashing algorithm is used that projects the input values onto the hash code value range in as close a random distribution as possible, and a sufficiently large bit length is chosen, then the dataset to be hashed can be encoded with a statistically extremely high probability that no two different structures will result in the same hash code ("collision"). The "price" for these desirable properties is that a hash code is usually one-way, i.e., it is not intended to allow the reconstruction of the input structure from the hash code value. A vast body of literature exists in the information and computer sciences on hashing and how to design and code fast and reliable algorithms. A hash code of 64 bit length can be shown to be sufficient to encode even the largest existing chemical dataset such as the CAS Registry with negligible risk of collision.

A useful hash code of chemical structures has to possess the additional property that it generates the same hash code for compounds that are chemically identical but differently coded in the input representation used. For example, it cannot simply use the connection table with its arbitrary atom numbering as the input value. Ideally, invariant molecule properties are used to generate the seeds for the computation of the hash codes (although canonicalzation of the connection table has also been used).

A reliable hash code of chemical structures that fulfils the above criteria is then ideally suited for the task of analyzing internal duplication within, and overlap between, large chemical databases. The CACTVS system[33,34] generates such hash codes and was therefore used for these types of analyses.

The CACTVS system version 3.00 was used to calculate the nonstereospecific and stereospecific perturbed[35] hash codes (ensemble properties E_HASHY and E_HASHSY). Deuterium isotopes were taken into account by appending to the hash code the molecular formula including isotope labels. These hash codes are 64 bit long, thereby ensuring collision-free encoding of the total of the compounds analyzed in this study (on the order of a million). Nonstereospecific hashing in CACTVS yields the same hash code for compounds with the same constitution. Stereospecific hashing produces three different hash codes if a substructure element is present that is "stereocritical," i.e., a chiral center (R/S) or a double bond with E/Z isomerism. If the stereochemistry is specified, one hash code will result for the R (or E) isomer and a different one for the S (or Z) isomer. In case the chirality or double bond geometry is not specified, the nonstereospecific hash code results.

In the same script that generated the hash codes, the number of duplicates in each database was determined, and thus a list of the unique compounds of each database was obtained. A PERL script allowed comparison of each of these eight lists with each other, yielding, for each unique compound (represented by the hash code), a list of the databases which contain this compound.

Traditionally, and up to this day, no stereochemistry information is encoded in any structure entered in the NCI database.[36] Therefore, no 3D coordinates were included in the generated SD file, which caused E_HASHY and E_HASHSY to be identical for each compound in the NCI database. When comparing the NCI database with the other databases (which all contain at least some stereochemistry information) using stereospecific hash codes, this leads to the possibility that, for stereocritical molecules, structures were not equated that are in reality identical. This stems from the fact that for the NCI database, each compound received a nonstereospecific hash code ("stereoinformation not specified"), whereas in the other database(s) the same stereoisomer may be present but with a stereospecific hash code, which is different from the nonstereospecific one even if the same 3D structure is hashed. To assess the magnitude of this error rate, the number of stereocritical compounds in each database was determined by calculating E_HASHY and E_HASHSY using an SD file with coordinates generated by the 2D → 3D conversion program CORINA,[37] which assigns a default stereochemistry to undefined centers. If E_HASHY and E_HASHSY are identical, the compound has no stereocenters and double bond isomers, otherwise the compound is stereocritical. If was found that 42.6% of the open NCI compounds are stereocritical.

Descriptors. Three different sets of descriptors were compared to each other in performance, and two (Ghose-Crippen and E_SCREEN) were used in the clustering analyses conducted in this study.

The CACTVS system implements a 431 bit fingerprint descriptor (E_SCREEN), which comprises 86 bits for binned element counts, 64 bits referring to absence/presence of atom pairs, 148 bits representing various ring sizes and types, and 133 bits representing larger fragments.

Descriptors based on the alogP atom type[38] have recently been successfully deployed in two publications[39,40] to determine if a molecule has drug-like properties or not. Adopting the approach used by these authors, we calculated these Ghose-Crippen atom type counts using a C program written by one of these authors (M. Wagener, who made this program available to us). A binary binning scheme was devised in which every bin was defined by a power of two (e.g. 1, 2, 4, 8, ...). For a given atom type count, all bits up to this number were set. For example, for the atom type of "count 9," the bits for bin 1, bin 2, bin 4, and bin 8 were set. The highest number for each atom type was determined by taking the highest atom type count in the NCI, ACD, and WDI databases combined. This resulted in a 494-bit vector. This binning procedure was realized with a PERL script.

In recent work comparing various descriptors,[41,42] the 166 "public" MDL keys, which represent small topological substructure fragments, were found to perform best for distinguishing between active and inactive compounds and for the prediction of physicochemical properties by various clustering methods. For this study, MDL keys were obtained using the ISIS system (MDL Information Systems Inc., San Leandro, CA), yielding a 167-bit vector.

Descriptor Assessment. To compare the three descriptors to each other by analyzing, for each, the correlations they produce between structure and biological activity, a method published by Bajorath[20] was employed. We used the same seven activity classes as were used in this publication but selected a different set of compounds for our test. The 705 compounds chosen were taken from the WDI because, for this database, the property field for activity class or mechanism of action allows assignment of each compound to a specific activity class (see Table 2). All counterions were removed (see remark below). The pairwise similarity between two fingerprint descriptors was calculated employing the Tanimoto coefficient,[43] defined as $N_{AB}/(N_A + N_B - N_{AB})$, with N_{AB} the number of common bits and N_A and N_B the numbers of bits set in A and in B, respectively. For each molecule in the test database, the neighbors within a given similarity cutoff were considered, and the percentages of correctly ("*C*," same activity class as compound) and incorrectly ("*I*," different activity class than compound) assigned molecules were calculated. For each activity class, *C* and *I* were summed up, and an activity class score *S* was calculated according to $S = (\sum C - \sum I)/N$, where *N* is the total number of compounds in this activity class. If $\sum C$ was smaller than $\sum I$, *S* was set to zero. The total score S_{total} was the average over all seven activity classes. A PERL script accomplished this assessment procedure.

Table 2. Test Database for Descriptor Assessment

activity class	no. of compds
H3 antagonists (H3)	31
carbonic anhydrase II (CA) inh.	54
HIV protease inh. (HIV)	134
serotoninergic (SER)	145
benzodiazepine receptor ligands (BEN)	232
cyclooxygenase-2 (COX) inh.	34
tyrosine kinase (TK) inh.	121

Similarity Overlap. The similarity overlap between two databases was determined following an approach by McGregor et al.[19] For each compound in the first database, the similarity (Tanimoto coefficient) to all compounds in the second database was computed, yielding the percentage of compounds of the first database for which a compound exists in the second database with a similarity greater or equal to a specified cutoff. This algorithm was implemented in a C program using the fast function for the Tanimoto coefficient computation as mentioned below.

Diversity Analysis by Clustering. To estimate the diversity of the databases, a clustering approach was chosen. The ideal clustering algorithm should yield, at a given similarity cutoff, the most representative subset with the least redundancy. Tominaga[44] compared nine different subset selection methods by box counting analysis. He found that the maximum dissimilarity selection algorithm, which was first proposed by Maggiora,[45] was strong at yielding subsets which were diverse and representative of the overall composition of the database. The basic concept of this algorithm is that each compound added to the subset must be most dissimilar to all compounds in the subset already chosen. Among the many methods based on this idea,[46,47] Nilakatan et al. deployed their DIVPICK program[18] to assess the diversity of databases.

The algorithm we used (see Scheme 1) is based on the optimal dissimilarity selection method (ODS) proposed by Clark.[48] It uses dissimilarity comparison of already selected compounds with subsamples of an adjustable size *K* of the entire remaining (unselected) database of size *N* ($K \leq N$), discarding subsample compounds that are more

Scheme 1. Explanation of the Optimal Dissimilarity Selection Method (ODS)

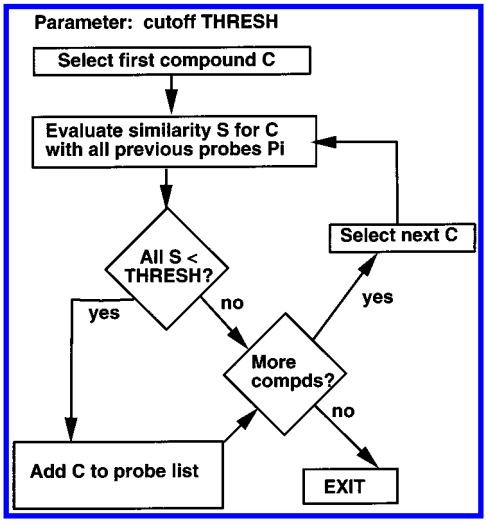
Parameters and sets: -similarity cutoff **THRESH**
- evaluation sample size **K**
- database: initial **POOL**
- subset **SUB**

Initialization:
calculate all the pairwise similarities in **POOL**, add the two compounds with lowest similarity to **SUB** and remove from **POOL**

Iteration:
1. Step: take compounds from **POOL*** discarding and removing all compounds which have similarity > **THRESH** to any compound of **SUB** until **K** compounds are chosen
2. Step: add to **SUB** and remove from **POOL** the compound with the lowest sum of similarities to all members in **SUB**
3. Step: leave remaining compounds in **POOL** and go to 1. Step

*: Procedure to get **POOL** compounds during iteration:
- start with first **POOL** compound
- if end of **POOL** is reached go to first **POOL** compound
- stop when **POOL** is empty

Scheme 2. Dissimilarity Step of the Stochastic Clustering Analysis (SCA)[49]



similar to the selected compound than a specified similarity cutoff. For larger databases it is impossible to iteratively compare all remaining compounds of the database to all the compounds already selected for the subset, i.e., one has to choose $K \ll N$. Therefore, we felt that, *vis-à-vis* the database sizes encountered in this study, i.e., with N ranging from ~50 000 to ~250 000, a subsample size of $K = 1000$, while being far from the "maximum dissimilarity" mode ($K = N$), would preserve a reasonable balance between diversity and representativeness, while still being computationally feasible.

Nevertheless, the ODS type method can take up to a week for databases with 250 000 compounds and therefore faster methods are desirable. As one example of these, diverse subsets can be computed using the dissimilarity step of the Stochastic Clustering Analysis algorithm (SCA) described by Reynolds et al.[49] (This method is essentially equivalent to ODS with $K = 1$.) The dissimilarity step of the SCA calculates a subset of the input vectors generated from a database such that each vector in the subset is dissimilar enough to all other vectors in the subset (see Scheme 2). Thus the selected vectors are evenly scattered over the whole space of the database. The vectors selected during the dissimilarity step are used as cluster centers to which the remaining structures are assigned. This assignment step was not performed in this work because cluster memberships are not needed. Also unlike the published algorithm, the order of input of the database vectors was not randomized but used as saved in the input file. The size of the subset (which is equivalent to the number of clusters) obtained is not predetermined. It is dictated by the molecular diversity of the database and the similarity cutoff. Reynolds et al.[49] suggest that the number of selected vectors obtained for a given similarity threshold may characterize the diversity of a database.

Table 3. Computation Times vs Tanimoto Coefficients to Calculate Subsets of the NCI Database (250 K Entries) Using Stochastic Cluster Analysis (SCA) on a 500 MHz Linux Computer

Tanimoto coeff	subset size	time (s)
0.1	17	29
0.2	39	29
0.3	100	29
0.4	260	29
0.5	777	33
0.6	2933	57
0.7	10911	156
0.8	35471	1845
0.9	94594	8223

Table 4. Number and Percentage of Compounds Which Are Present in One Database Only

nonstereospecific hash code			stereospecific hash code		
ASINEX	126782	92.0%	ASINEX	127921	92.8%
CSD	64081	88.2%	CSD	67965	91.2%
NCI	194160	82.5%	NCI	201282	85.5%
WDI	41147	76.3%	WDI	46774	83.4%
SIGALCAT	6847	75.2%	ACD	114043	52.5%
ACX	5269	53.6%	MAY	15328	27.7%
ACD	79362	37.6%	SIGALCAT	13045	14.0%
MAY	9250	16.8%	ACX	12535	12.7%

Since one of our goals was to compare different databases, we express the diversity of a database by dividing the number of clusters obtained by both ODS and SCA by the number of entries in the database. Calculation of the diversity percentages for the E_SCREEN and the Ghose-Crippen descriptors at similarity cutoff of 0.80 and 0.68, respectively, showed a virtually identical ranking (see Table 10). A Tanimoto similarity cutoff of 0.80 for E_SCREEN and a Tanimoto similarity cutoff of 0.68 for Ghose-Crippen produced the highest score in the distinction of biological activity by structure, as the results in Figure 1 show. Therefore usage of the much faster SCA clustering for the analysis of varying similarity cutoffs was deemed to be justified.

Both algorithms were implemented in C programs we called SUBSET (for SCA) and OPTISIM (for ODS).[50] Unlike many other scientific programs, which have arbitrary limitations for array sizes etc., SUBSET and OPTISIM have no limits for the number of entries, the subset size, and the length of the bit strings. All data structures grow dynamically when needed, up to the size of the available physical memory. Great care was taken to ensure C code correctness by observing good coding practices and applying code validation tools. We believe that the algorithm used in SUBSET is efficient and capable of tackling large databases. In our implementation, calculating the Tanimoto coefficient between two 431 bits vectors can be done about 1 900 000 times per second on an Intel 500 MHz Pentium II Linux computer. On the same machine, calculation of a subset of 34 471 vectors from the NCI database (~250k entries) using 431 bits vectors and a typical Tanimoto coefficient of 80%

Table 5. Overlap of Identical Compounds (Nonstereospecific Hash Codes) in Absolute Numbers and Percent[a] of Row Database

	ASINEX	ACX	SIGAL-CAT	MAY	NCI	CSD	WDI	ACD
ASINEX	137810	6089 (4.4%)	5974 (4.3%)	1201 (0.9%)	3690 (2.7%)	671 (0.5%)	410 (0.3%)	7384 (5.4%)
ACX	6089 (6.2%)	98318	79576 (80.9%)	5379 (5.5%)	25719 (26.2%)	4032 (4.1%)	5352 (5.4%)	81427 (82.8%)
SIGALCAT	5974 (6.6%)	79576 (87.4%)	91072	4488 (4.9%)	23760 (26.1%)	3837 (4.2%)	4625 (5.1%)	72316 (79.4%)
MAY	1201 (2.2%)	5379 (9.8%)	4488 (8.2%)	55064	5416 (9.8%)	811 (1.5%)	496 (0.9%)	45084 (81.9%)
NCI	3690 (1.6%)	25719 (10.9%)	23760 (10.1%)	5416 (2.3%)	235461	5302 (2.3%)	7793 (3.3%)	31331 (13.3%)
CSD	671 (0.9%)	4032 (5.5%)	3837 (5.3%)	811 (1.1%)	5302 (7.3%)	72680	2710 (3.7%)	4721 (6.5%)
WDI	410 (0.8%)	5352 (9.9%)	4625 (8.6%)	496 (0.9%)	7793 (14.5%)	2710 (5.0%)	53895	7148 (13.3%)
ACD	7384 (3.5%)	81427 (38.5%)	72316 (34.2%)	45084 (21.3%)	31331 (14.8%)	4721 (2.2%)	7148 (3.4%)	211274

[a] The percentages are calculated as the fraction of the overlap sets relative to the row database. For example, the leftmost value in the bottom row means that 3.5% of the ACD is also present in ASINEX.

Table 6. Overlap of Identical Compounds (Stereospecific Hash Codes) in Absolute Numbers and Percent[a] of Row Database

	ASINEX	ACX	SIGAL-CAT	MAY	NCI	CSD	WDI	ACD
ASINEX	137835	5646 (4.1%)	5392 (3.9%)	1128 (0.8%)	3010 (2.2%)	519 (0.4%)	341 (0.2%)	5230 (3.8%)
ACX	5646 (5.7%)	98788	74016 (74.9%)	4973 (5.0%)	23812 (24.1%)	2973 (3.0%)	3964 (4.0%)	58405 (59.1%)
SIGALCAT	5392 (5.8%)	74016 (79.4%)	93243	4084 (4.4%)	21090 (22.6%)	2883 (3.1%)	3415 (3.7%)	53627 (57.5%)
MAY	1128 (2.0%)	4973 (9.0%)	4084 (7.4%)	55262	4345 (7.9%)	653 (1.2%)	403 (0.7%)	38307 (69.3%)
NCI	3010 (1.3%)	23812 (10.1%)	21090 (9.0%)	4345 (1.8%)	235461	3157 (1.3%)	4937 (2.1%)	21905 (9.3%)
CSD	519 (0.7%)	2973 (4.0%)	2883 (3.9%)	653 (0.9%)	3157 (4.2%)	74510	1772 (2.4%)	4341 (5.8%)
WDI	341 (0.6%)	3964 (7.1%)	3415 (6.1%)	403 (0.7%)	4937 (8.8%)	1772 (3.2%)	56096	4994 (8.9%)
ACD	5230 (2.4%)	58405 (26.9%)	53627 (24.7%)	38307 (17.6%)	21905 (10.1%)	4341 (2.0%)	4994 (2.3%)	217216

[a] The percentages are calculated as the fraction of the overlap sets relative to the row database. For example, the leftmost value in the bottom row means that 2.4% of the ACD is also present in ASINEX.

Table 7. Number of Compounds Which Are in Exactly/At Least *n* Databases (DBs)

	nonstereospecific hash code		stereospecific hash code	
no. of DBs *n*	no. of compds in at least *n* DBs	no. of compds in exactly *n* DBs	no. of compds in at least *n* DBs	no. of compds in exactly *n* DBs
1	681157	526898	737874	598893
2	154259	74294	138981	76476
3	79965	49465	62505	40160
4	30500	22852	22345	16970
5	7648	5832	5375	4182
6	1816	1594	1193	1058
7	222	215	135	132
8	7	7	3	3

Table 8. Similarity Overlap in Percent[a] between Databases[b]

	ASINEX	ACX	SIGALCAT	MAY	NCI	CSD	WDI	ACD
ASINEX	100	84	87	78	90	64	58	92
ACX	74	100	99	77	94	83	76	97
SIGALCAT	76	98	100	78	94	83	76	97
MAY	74	73	73	100	85	58	52	99
NCI	72	84	84	71	100	78	72	88
CSD	47	63	62	45	73	100	51	67
WDI	64	84	83	58	92	84	100	89
ACD	75	87	87	84	91	73	69	100

[a] Percentage of stuctures in row database which have a similarity match, at a Tanimoto coefficient ≥ 0.80, in column database. [b] E_SCREEN descriptor at Tanimoto coefficient ≥ 0.80.

takes about 30 min and 4 MB of memory. We found that computation times increase nonlinearly with the similarity coefficient as shown in Table 3.

Many of the structures analyzed in the context of this study contain one or more counterions, and one has to make the decision whether to include them in the individual analyses. For the calculation of both the Ghose-Crippen bitstrings and E_SCREEN bitstrings of the eight large databases the counterions were not removed. This is in contrast to the procedure applied to the small test database of 705 WDI

Table 9. Similarity Overlap in Percent[a] between Databases[b]

	ASINEX	ACX	SIGALCAT	MAY	NCI	CSD	WDI	ACD
ASINEX	100	89	89	83	93	72	67	94
ACX	76	100	99	80	97	87	82	99
SIGALCAT	79	99	100	81	97	87	81	99
MAY	78	79	80	100	87	63	59	100
NCI	74	88	87	73	100	81	76	92
CSD	58	79	78	54	86	100	63	81
WDI	62	86	85	55	93	84	100	90
ACD	77	90	90	85	92	76	73	100

[a] Percentage of stuctures in row database which have a similarity match, at a Tanimoto coefficient ≥ 0.68, in column database. [b] Ghose-Crippen descriptor at Tanimoto coefficient ≥ 0.68.

used for the descriptor assessment, where we did remove the counterions prior to descriptor calculation. The reason is that, in the latter case, we were primarily interested in the diversity of the databases with respect to biological activity, for which generally it makes sense to remove counterions, whereas for the large database analysis we were interested in the chemical diversity in general.

Hardware. All calculations of similarity overlaps and clustering were performed on an Intel 500 MHz Pentium II Linux computer. All other calculations were done on an SGI Octane.

RESULTS AND DISCUSSION

Analysis and Comparison of the Databases. Table 1 gives an overview over the databases compared in this study, the number of nonduplicate structures with stereochemistry taken into account or not, and the presence or absence of stereochemistry information in each database. Also, a very approximate price range is given.

As can be seen, the Open NCI Database is the largest dataset in this comparison with nearly 250 000 structures, followed by the ACD with about 222 000 in the version analyzed. All other databases were significantly smaller.

Table 10. Diversity Ranking (Number of Clusters at Given Similarity Cutoff Divided by the Number of Compounds in Database) for Optimal Dissimilarity Selection (ODS) and Stochastic Cluster Analysis (SCA)

E_SCREEN (Tanimoto coeff. ≥ 0.80)				Ghose−Crippen (Tanimoto coeff. ≥ 0.68)			
ODS	%	SCA	%	ODS	%	SCA	%
CSD	39.1	CSD	36.2	CSD	30.2	CSD	28.0
MAY	23.0	MAY	20.4	MAY	22.3	MAY	20.2
WDI	20.2	WDI	17.8	WDI	19.8	WDI	18.1
SIGALCAT	17.0	SIGALCAT	15.1	NCI	15.44	SIGALCAT	13.7
NCI	16.5	NCI	14.2	SIGALCAT	15.40	NCI	13.6
ACX	15.7	ACX	13.8	ACX	14.1	ACX	12.6
ACD	14.1	ACD	12.2	ACD	13.7	ACD	12.3
ASINEX	10.0	ASINEX	8.6	ASINEX	9.6	ASINEX	8.8

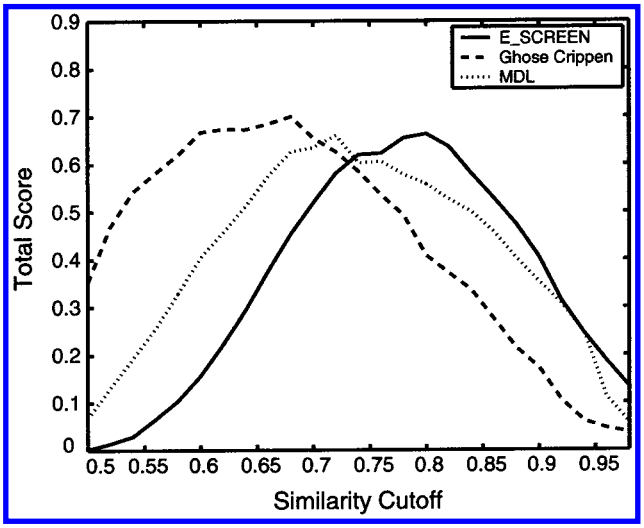


Figure 1. Score S_{total} for descriptors at different similarity cutoffs.

Many of the databases, in particular the commercial collections, have entries that have no structure associated with them, thus reducing—sometimes significantly (SIGALCAT)—the numbers of entries that were relevant for our study. Presumably, such entries are products, that, e.g., are mixtures, polymers, items such as silica gel and indicator sets, etc., that cannot be represented by a well-defined structure. One should keep this aspect in mind when these collections are assessed from a structure database point of view. Since only the 3D version of the ACD was considered, in which obviously all compounds have a structure, no precise number of entries without a structure as a percentage of the entire ACD can be reported (∼278k unique chemicals claimed).

Apart from structureless entries, a nonzero duplication rate was found within each of the analyzed databases. If stereospecific hash codes were used, this internal duplication rate ranged from 0.02% (MAY) to 13.0% (ACX). This method is obviously the more appropriate approach for those databases that have stereochemistry information—both chirality and E/Z stereochemistry—consistently associated with all of their structures. However, in cases where stereochemistry indicators are not applied correctly throughout the entire database (e.g. in the ACD), this may actually determine the uniqueness rate of the database incorrectly. The NCI database, having no stereochemistry whatsoever associated with its structures, does not allow one to distinguish between R/S and E/Z isomers. In other words, this uniqueness analysis has, for the NCI database, declared many structures as identical—and therefore nonunique—that are in reality different stereoisomers. (This is known from the fact that, for that subset of the NCI database for which names are present in the original tables, these names often do give an indication that one is facing different stereoisomers—information that, unfortunately, has not been easy to evaluate automatically in a reliable manner; however, efforts are underway to tackle this problem.) To allow a comparison of the other databases with the NCI database in this regard, this "worst case scenario" uniqueness analysis based on nonstereospecific hash codes was also applied to all the other databases.

Still, even for the NCI database, there is a good deal of internal duplication present that is not caused by the lack of stereochemistry information, as some anecdotal analysis showed. (However, we cannot give an accurate quantitative ratio of these cases vs the false stereochemistry identity cases exactly because stereochemistry information is lacking.) There are instances where the same chemical was submitted again—perhaps many years later—by possibly a different supplier for testing by NCI and thus received a different NSC number. More common are cases where the same organic molecule was part of a different formulation, which can be gleaned from the text records (including the name fields) in the original file, but could not be, and thus had not been, expressed in the structure record in the original NCI database entry. Even more common appears to be the case where the original chemical submitted was a structure that would be represented by two disconnected fragments, as in the case of metal salts, but for which one fragment, typically a metal ion, did not "survive" the translation process of the original data into to the .mol files provided by DTP. As a consequence, the hash code, formed without this additional information, rendered all these entries structurally identical.

Two examples may illustrate these cases. Nitrate (HNO_3) was the most common structure found in the NCI database, occurring 37 times. A look at the individual entries however showed that these are in reality salts of different metal ions, such as gallium nitrate (NSC# 15200), indium nitrate (NSC# 15202), cesium nitrate (NSC# 84271), etc. The second most common molecule found by hash code in the NCI database (28 instances)[51] was daunomycin (in fact known under many other names), with the molecular formula $C_{27}H_{29}NO_{10}$, identified without stereochemistry information by the IUPAC name of 3-acetyl-3,5,12-trihydroxy-10-methoxy-6,11-dioxo-1,2,3,4,6,11-hexahydro-1-naphthacenyl 3-amino-2,3,6-trideoxy-hexopyranoside. In this case, analysis of the NCI-supplied names showed that one is mostly dealing both with instances of different stereoisomers and with mixtures that could not be expressed in a structure, such as "salmon sperm DNA complex" (NSC# 257454), "mixture with denatured calf thymus DNA" (NSC# 262200), etc. Nine (different) CAS RNs are present in the set of 28 entries. The number of names

stored in the original record of each of the 28 NSC numbers ranges from zero to 20 (for NSC# 82151), with four entries having no name associated with them (but all four do have a CAS RN).

The Open NCI Database is about evenly split between entries that have a CAS RN associated with them (126 830 compounds, 50.9%) and those that do not (122 251 compounds, 49.1%). This has to do mostly with the history of the keying-in of the original data. A preliminary analysis seemed to indicate that there are, in fact, quite a few compounds in the NCI database that truly do not have a CAS RN, presumably because they were never published in a journal article or patent. Further analyses on this will be reported in the future.

For the other databases, especially the collection of commercially available chemicals, the internal duplication is typically caused by different amount units, purity grades, formulations, etc. Unresolved overlap between vendor catalogs and/or weaknesses in the structure identification algorithm used to build the database may additionally contribute to intradatabase overlap.

The ACX stands out with the highest internal duplication rate using either algorithm. No detailed analysis of this situation was conducted. The SIGALCAT does not fare much better when the number of entries is compared with the number of unique structures. However, for this database, the main reduction in numbers occurred because of structureless entries. For the CSD, the internal duplication is, in all likelihood, mainly due to multiple entries of crystals of the same structure by different authors or different crystallization or crystallography conditions in entries by the same author(s).

Another aspect of uniqueness is the question of how many compounds, for each database, occur *only* in that database. Table 4 lists the answers, again calculated both with nonstereospecific and with stereospecific hash codes. Table 4 also lists, and is sorted by, the fraction that the subset of such unique compounds constitutes as a percentage of the entire database.

While ASINEX and CSD lead in terms of their fraction of unique compounds within this comparison, the NCI database has by far the highest absolute number of unique structures. On the order of 200 000 compounds in the NCI database are present only there and do not occur in any other of the seven analyzed databases. This is explained by, and reconfirms, our prior finding[7] that the NCI database is a very eclectic collection of structures with many unusual compounds in it that do not fall under the scope of commercially available chemicals, standard drugs, or crystallographically studied structures. Whether this is an advantage or a disadvantage of the NCI database obviously depends on the purpose for which one wants to use it.

Generally, with the exception of ASINEX, the databases of commercially available compounds (ACD, MAY, SIGALCAT, ACX) show the lowest fractions of unique compounds. This is not surprising given their nature of catalogs of many commercially available compounds, for which a large degree of mutual overlap is to be expected. Still, not a single database was entirely a subset of any of the other databases, showing that each one of them may have its own specialty area and thus value for the user.

The sizes of the overlap sets—both in absolute numbers and expressed in percentage of the database—for each pair of databases are listed in Table 5 for nonstereospecific hash codes and in Table 6 for stereospecific hash codes, respectively. Concentrating on the overlaps calculated with stereospecific hash codes, which are probably the somewhat more relevant numbers for most of the databases, we see that the single largest overlap, both in absolute numbers and percentages, occurs between ACX and SIGALCAT; that the ACD contains a good percentage—between 57% and 70%—of each of the other three commercial catalogs; that the smallest average overlap (2.4%) is observed for the CSD; that the overlap sets of the NCI database range from 3010 (1.3%) with ASINEX to 23 812 (10.1%) with the ACX; and that the overlap of the NCI database with the ACX is also the largest when taken as the fraction of the other database (to be found by looking at the *columns* in Tables 5 and 6), in this case 24.1% of the ACX. These, and other, basic findings do not change much if one studies the nonstereospecific version of the overlap analysis.

With regard to the commercial compound databases one may want to note that ACD and ACX both claim to cover the complete Sigma-Aldrich catalog. Therefore the overlap of only 57.5% (stereospecific hash codes) of the SIGALCAT compounds in the ACD is somewhat surprising. The ACD also should cover the entire MAY. However, using the stereospecific hash codes, 69.3% of MAY were found to be in the ACD. Since CamSoft distributes ACX and MAY separately, the overlap of only 9.0% of the MAY compounds in the ACX is understandable. Reasons for the discrepancy between claimed 100% overlap and the lower observed overlap percentages might be use of different database versions, errors in the structural representations (e.g. missing charges in SIGALCAT), or problems with missing labeling of racemic compounds and compounds with undefined stereochemistry (e.g. ACD).

The pairwise overlaps are obviously one type of intersection that can be studied. One could additionally ask, e.g., what the size of the set is of all compounds that occur in all three of the noncommercial databases, NCI, CSD, and WDI, but not in any of the remaining databases, and so forth. The number of all such possible combinations of questions, and thus answers, for eight databases is $2^8 - 1 = 255$. It is clearly neither practical nor probably of general interest to include all these values here. It is, however, very easy to derive any of them, if so desired, with the software we used for this study.

Two sets of such additional overlap results that may be of particular interest shall be given here, though. First, a number regarding the NCI database that both has been of interest to us in our drug development work and was asked from us by users of the NCI database is the percentage of (open) NCI database compounds that are also available commercially from any source. While we cannot claim that our analysis has included the catalogs of every chemicals' supplier in the world, the collections evaluated here are assumed to cover a significant percentage of the universe of commercially available compounds. Analyzing the overlap of the NCI database with the union of the ACD, ACX, ASINEX, MAY, and SIGALCAT collections, we found that, based on nonstereospecific hash codes, 35 545 NCI compounds (15.1%) are commercially available, while this number is 30 945 (13.1%) when stereospecific hash codes are used.

Second, it was of interest to us to know how many compounds are in exactly *n*, or, respectively, in at least *n*, databases, *n* ranging from 1 to 8. Again, this analysis was performed using both nonstereospecific and stereospecific hash codes. Table 7 lists the results, giving both the individual set sizes for each value of *n* (compounds that are exactly in *n* databases) and the cumulative set sizes (compounds that are in at least *n* databases). The stereospecific analysis yielded that there are only three compounds that are found in each of the eight databases studied. At the other end of the range, the number of stereochemically different compounds that are in at least one database was 737 874, which thus constitutes the entire universe of different chemical structures present in this study. The simple sum of the individual database sizes (stereospecific, internally unique hash codes; see Table 1) yielded 968 411, whereas the total number of structures across all databases, disregarding database-internal duplication, was 1 016 995.

Descriptor Assessment. The results shown in Figure 1 demonstrate that all three descriptors performed quite similarly with the test set chosen. While the Ghose-Crippen descriptors yielded a slightly higher S_{total} score than the other two, that difference is probably not statistically relevant. Likewise, the overall curve shapes are very similar, showing roughly Gaussian distribution. The most pronounced difference between the three descriptors is the similarity cutoff value at which each curve reaches its maximum, which was approximately 0.68 for Ghose-Crippen, approximately 0.72 for the MDL keys, and approximately 0.8 for the E_SCREEN fingerprint descriptors. Either one of these optimal similarity cutoffs yielded a comparable number of clusters with either clustering method (see Table 10) and thus appear to be generally comparable in terms of the degree of similarity they produce.

A tentative explanation of the different cutoff maxima for the three descriptors are the different ways structural information is encoded. While the Ghose-Crippen bits are a cumulative binned representation of 121 atom type counts, the MDL keys directly represent absence/presence of 167 particular fragments. Therefore, if two compounds differ in the presence of one fragment, they differ only in one bit in the case of the MDL keys, whereas for the Ghose-Crippen bits this changes multiple bits by altering the respective atom counts for all the atoms comprising this particular fragment. E_SCREEN is a combination of cumulative binned representation of element counts, ring counts, and fragment bits (46% of the bits). Since this combination leads to a certain redundancy (cross-correlation of bits), the optimum similarity cutoff for this descriptor would naturally expected to be the highest.

One would assume that any well-designed descriptor set that is capable of capturing the vast majority of the structural and/or physicochemical variance within a given compound set should be able to deliver results that obey the "similar property principle,"[52] which assumes that structurally similar compounds also have similar biological and physicochemical properties. Our results indicate that all three descriptors fulfill this condition, therefore leading to the freedom to choose the descriptor type that may be the most convenient one for other reasons.

Similarity Overlap. Tables 8 and 9 show the results of the similarity overlap analysis between all databases for the E_SCREEN and the Ghose-Crippen descriptors, respectively. As was to be expected from the similarity of the curves in Figure 1, the two descriptors, adjusted for their differing optimal cutoffs, delivered quite similar results for the similarity overlaps. In practically all cases, at least half of the structures in any of the databases had a similarity match in each of the other databases; typically, the overlap percentage was more likely to be in the 70−99% range. The commercial databases showed similarity overlaps to each other that were quite comparable, approaching 100% in several cases. One would assume that the larger a database is, the more likely it is to find in it a structure similar to a given molecule (from another database). This was borne out by the fact the two largest databases, ACD and NCI, show generally the largest similarity percentages in Tables 8 and 9. The lowest similarity percentages were found for WDI, MAY, and CSD, respectively, again being readily explainable by the fact that these were the smallest databases in the set.

Clustering. All four combinations of using the two different descriptors (E_SCREEN and Ghose-Crippen) and applying both clustering methods (ODS and SCA) to each of them were evaluated by tabulating the numbers of clusters at each descriptor's optimum similarity cutoff divided by the number of compounds in each database (Table 10). Generally, these values were slightly lower for Ghose-Crippen than for E_SCREEN and, likewise, slightly lower for SCA when compared to ODS. Nevertheless, they yielded virtually the same diversity ranking of the databases.

More pronounced than the differences between descriptors/methods were the differences between databases. They fell into four groups according to their similarity percentages, with the CSD ranking at the top by itself, close to or above the 30% mark; MAY and WDI were around 20%, within approximately 3% points of each other; SIGALCAT, NCI, ACX, and ACD were typically around the 15% range, and within 3% points of each other; ASINEX was at or below the 10% mark in all cases. While the size of the database may have some bearing on these results, too (the largest ones, NCI and ACD being in the bottom half), the third largest one, ASINEX, found itself at the very bottom of the diversity ranking. The three compilations of different commercial catalogs, ACX, ACD, and ASINEX, are uniformly lowest in diversity, which is not too surprising given the fact that one can expect a substantial similarity overlap between the various manufacturers of chemicals. Conversely, the CSD, having no unifying structural theme and presumably the least ties to commercial sources of all eight databases would indeed be expected to have the widest diversity of structural motifs in it.

When one plots the clusters/entries ratio as a function of the similarity cutoff for both the E_SCREEN (Figure 2) and the Ghose-Crippen (Figure 3) descriptor, reasonably similar trends emerge. For either descriptor, the CSD was ahead of all other databases except at very high cutoff values, where it was surpassed by MAY. For both descriptors, the databases fell into two groups when categorized by the steepness of their curves. MAY, NCI, ACD, and ASINEX all showed a steeper slope at higher similarity cutoff values ($\gtrsim 0.8$) than the other four databases, which may mean that there are substantial subsets of quite similar structures in them, which only get resolved into different clusters when high cutoff values are applied.

Table 11. Overview over Characteristics and Statistics of the Open NCI Database (October 1999 Version)

no. of entries	249 081 (100%)
no. of nonduplicate entries	235 461 (94.5%)
no. of entries not present in any other analyzed DB (nonstereospecific)	194 160 (82.5%)
no. of entries not present in any other analyzed DB (stereospecific)	201 282 (85.5%)
no. of entries with CAS Registry Number	126 830 (50.9%)
no. of entries that are stereocritical (R/S or E/Z)	106 122 (42.6%)
no. of entries with names in the original DTP data	45 228 (18.1%)
overlap with union set of commercial DBs[a] (nonstereospecific)	35 545 (15.1%)
overlap with union set of commercial DBs[a] (stereospecific)	30 945 (13.1%)
overlap with World Drug Index (nonstereospecific)	7793 (3.3%)
overlap with World Drug Index (stereospecific)	4937 (2.1%)
overlap with Cambridge Structural Database (nonstereospecific)	5302 (2.3%)
overlap with Cambridge Structural Database (stereospecific)	3157 (1.3%)
price	free

[a] ACD, ACX, ASINEX, MAY, SIGALCAT.

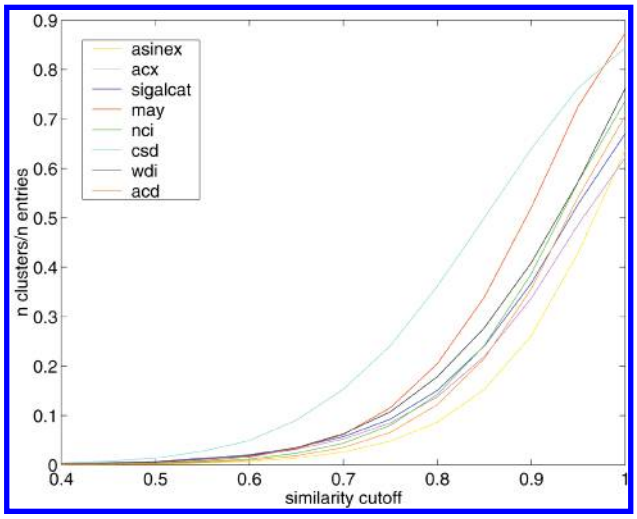


Figure 2. Number of clusters divided by the number of database entries vs similarity cutoff (E_SCREEN descriptor).

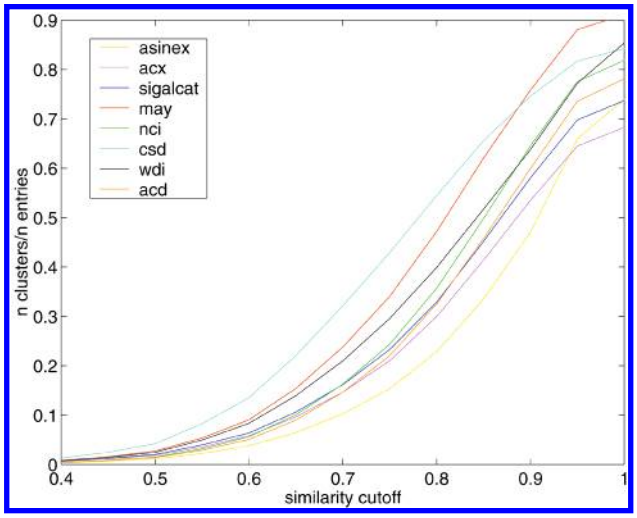


Figure 3. Number of clusters divided by the number of database entries vs similarity cutoff (Ghose-Crippen descriptor).

CONCLUSION

Each of the eight large chemical databases presented here has at least some overlap with each of the remaining ones. For some of the compilations of commercial vendor catalogs, this mutual overlap can be quite substantial (up to 80% of the smaller of the two databases). Still, even in instances where one such commercial compilation might be thought to be a total subset of another one, by virtue of their respective sizes, this was not the case a single time. Each database had at least 5000 unique structures even when analyzed nonstereospecifically (which might equate different stereoisomers). This means that each of these databases seems to have its own "niche" and that the user, looking for a variety of structures, may want to give consideration to each of them. The crystallographic database CSD stands somewhat apart as can be seen from its various overlap and diversity measures, which is to be expected from its purpose and history when compared to the other databases. The WDI showed a surprisingly low degree of overlap (both by structure identity and similarity) with the other commercial databases, which may indicate that the chemical subuniverses of medicinal drugs and commercial compounds are quite separate.

The Open NCI database, while not as separate from the rest of the group as the WDI or CSD, showed by far the highest number of unique compounds. Approximately 200 000 of the NCI structures were not found in any of the other analyzed databases. Its sheer size obviously contributes to this fact, but its eclectic nature, brought about by its history and its mode of collection, should not be overlooked. The Open NCI Database would therefore appear to be a valuable resource to the chemical community not just because it is free. It offers to the chemist who is looking for a large collection of compounds, be it with concomitant sample availability or not, a diverse and oftentimes unique structure set. However, its shortcomings should not be overlooked, the main one of which is the total lack of stereochemistry information in the original records. Table 11 summarizes the various statistics for the Open NCI Database.

The detailed results of these database comparison necessarily represent a snapshot in time since all the analyzed databases evolve over the years. Before basing a decision of acquisition and/or usage of any of the databases presented here, the user should therefore investigate the current situation of the specific database(s) as to changes in size, scope, vendors included, etc. Still, we believe that the analyses and comparisons reported here may give a useful overview over these structure collections and, in particular, may help put the Open NCI Database in perspective relative to some of the other most important large chemical databases available.

ACKNOWLEDGMENT

We gratefully acknowledge the contribution of Dan Zaharevitz of the DTP, NCI, to this study, and to the

chemical information field in general, by putting up on the World Wide Web the data of the Open NCI Database. We thank Wolf-Dietrich Ihlenfeldt of the CCC, Erlangen, for help and many useful suggestions with his chemical information toolkit CACTVS. We thank Markus Wagener of N. V. Organon, NL, for making available to us his atom type counting program. We thank David Covell, CCR, NCI, for his contributions to this study.

REFERENCES AND NOTES

(1) Milne, G. W. A.; Miller, J. A. The NCI Drug Information System. 1. System Overview. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 154−159.
(2) Milne, G. W. A.; Feldman, A.; Miller, J. A.; Daly, G. P.; Hammel, M. J. The NCI Drug Information System. 2. DIS Pre-Registry. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 159−168.
(3) Milne, G. W. A.; Feldman, A.; Miller, J. A.; Daly, G. P. The NCI Drug Information System. 3. The DIS Chemistry Module. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 168−179.
(4) Milne, G. W. A.; Miller, J. A.; Hoover, J. R. The NCI Drug Information System. 4. Inventory and Shipping Modules. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 179−185.
(5) Zehnacker, M. T.; Brennan, R. H.; Milne, G. W. A.; Miller, J. A. The NCI Drug Information System. 5. DIS Biology Module. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 186−193.
(6) Zehnacker, M. T.; Brennan, R. H.; Milne, G. W. A.; Miller, J. A.; Hammel, M. J. The NCI Drug Information System. 6. System Maintenance. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 193−197.
(7) Milne, G. W. A.; Nicklaus, M. C.; Driscoll, J. S.; Wang, S.; Zaharevitz, D. National Cancer Institute Drug Information System 3D Database. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1219−1224.
(8) http://dtp.nci.nih.gov.
(9) Typically, a 2-year waiting period after submittal is imposed even for "open" structures before they are made available to the public.
(10) http://dtp.nci.nih.gov/docs/dtp_data.html.
(11) http://cactus.nci.nih.gov.
(12) http://dtp.nci.nih.gov/branches/dscb/repo_open.html. It should be noted that no quality testing whatsoever has been, and is, performed by NCI for any of the samples upon receiving them from their suppliers, which would be a near impossible, and prohibitively expensive, undertaking for a database of this size. Nor is any sample tested when it is shipped from the NCI repository. It is therefore advisable to conduct analyses of the samples as to purity, possible decomposition, etc. before their use in an assay.
(13) http://www.cas.org/casdb.html.
(14) http://www.infochem.de/viniti.htm.
(15) http://www.Beilstein.com/products/xfire/.
(16) Shemetulskis, N. E.; Dunbar, J. B.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the Diversity of a Corporate Database Using Chemical Database Clustering and Analysis. *J. Comput.-Aided. Mol. Design* **1995**, *9*, 407−416.
(17) Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. Molecular Diversity in Chemical Databases: Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Compounds. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750−763.
(18) Nilakantan, R.; Bauman, N.; Haraki, K. S. Database Diversity Assessment: New Ideas, Concepts, and Tools. *J. Comput.-Aided. Mol. Design* **1997**, *11*, 447−452.
(19) McGregor, M. J.; Pallai, P. V. Clustering of Large Databases of Compounds: Using the MDL "Keys" As Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.
(20) Xue, L.; Godden, J. W.; Bajorath, J. Database Searching for Compounds With Similar Biological Activity Using Short Binary Bit String Representations of Molecules. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 881−886.
(21) Bernard, P.; Golbraikh, A.; Kireev, D.; Chretien, J. R.; Rozhkova, N. Comparison of Chemical Databases: Analysis of Molecular Diversity With Self-Organising Maps (SOM). *Analusis* **1998**, *26*, 333−341.
(22) This is essentially the subset of (open) NCI compounds that have a CAS Registry Number associated with them.
(23) http://www2.ccc.uni-erlangen.de/software/cactvs/index.html.
(24) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical-Structure File Formats Used by Computer-Programs Developed at MOLECULAR DESIGN LIMITED. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.
(25) http://cactus.nci.nih.gov/ncidb2/download.html—please contact corresponding author if URL is not valid any more.
(26) http://www.mdli.com.
(27) http://www.camsoft.com.
(28) http://www.maybridge.com/html/home.htm.
(29) http://www.asinex.com/welcome.htm.
(30) http://www.sigma-aldrich.com.
(31) http://www.derwent.com/worlddrugindex/index.html.
(32) http://www.ccdc.cam.ac.uk/.
(33) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, S.; Sasaki, S. CACTVS: A Chemistry Algorithm Development Environment. In *Daijuukagakutouronkai Dainijuukai Kouzoukasseisoukan Shinpojiumu Kouenyoushishuu*; Machida, K., Nishioka, T., Eds.; Kyoto University Press: Kyoto, 1992.
(34) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, S.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach Toward Modularity and Flexibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109−116.
(35) Computation of "perturbed" hash codes is an option in CACTVS that guarantees different hash codes among sets of highly symmetrical, but different, molecules such as cage structures.
(36) Stereochemistry information may be encoded, in an indirect manner, in the name(s) entered in the database for those compounds that have names associated with them (about 18% of the open NCI database).
(37) http://www2.ccc.uni-erlangen.de/software/corina.
(38) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for 3-Dimensional Structure Directed Quantitative Structure−Activity Relationships 0.4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally-Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.
(39) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating Between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(40) Wagener, M.; Van Geerestein, V. J. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280−292.
(41) 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.
(42) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand−Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(43) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(44) Tominaga, Y. Data Structure Comparison Using Box Counting Analysis. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 867−875.
(45) Lajiness, M. S.; Johnson, M. A.; Maggiora, G. M. Implementing Drug Screen Programs Using Molecular Similarity Methods. In *QSAR: Quantitative Structure−Activity Relationship in Drug Design*; Fauchère, J. L., Ed.; Alan R. Liss, Inc.: New York, 1989.
(46) Holliday, J. D.; Ranade, S. S.; Willett, P. A Fast Algorithm for Selecting Sets of Dissimilar Molecules From Large Chemical Databases. *Quantum Struct.−Act. Relat.* **1995**, *14*, 501−506.
(47) Matter, H. Selecting Optimally Diverse Compounds From Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.
(48) Clark, R. D. OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181−1188.
(49) Reynolds, C. H.; Druker, R.; Pfahler, L. B. Lead Discovery Using Stochastic Cluster Analysis (SCA): A New Method for Clustering Structurally Similar Compounds. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 305−312.
(50) The Tanimoto coefficient algorithm is available upon request from the authors.
(51) NSC numbers 82151, 83142, 112758, 169533, 249333, 249334, 257453, 257454, 257457, 262198, 262199, 262200, 262647, 262648, 262649, 272705, 272706, 273442, 275272, 301724, 302648, 303826, 303827, 304684, 304685, 309694, 312627, and 359653.
(52) Johnson, M.; Maggiora, G. M. Concepts and Applications of Molecular Similarity; Wiley: New York, 1990.

CI000150T