New Diversity Calculations Algorithms Used for Compound Selection

Sergei V. Trepalin,*,† Vadim A. Gerasimenko,† Andrey V. Kozyukov,‡ Nikolay Ph. Savchuk,‡ and Andrey A. Ivaschenko‡

Institute of Physiologically Active Compounds, 142432 Chernogolovka, Moscow Region, Russia, and ChemDiv, Inc. 11575 Sorrento Valley Road, #210, San Diego, California 92121

Received June 30, 2001

Some modifications were introduced into the previously described Centroid diversity sorting algorithm, which uses cosine similarity metric. The modified algorithm is suitable for the work with large databases on personal computers. For example, for diversity sorting of the database with the size greater than a million of records, less than 9 h are required (Pentium III, 800 MHz). The problem of selecting new compounds into the existing collection is examined to reach the maximum diversity of the collection. The article describes the new algorithm for the selection of heterocyclic compounds.

INTRODUCTION

Presently, molecular similarity and diversity calculations are widely used in drug design, especially the high-throughput screening for biological activity, and in combinatorial libraries design. Similarity calculations along with other SAR methods are mainly used at the lead optimization stage and for the development of focused libraries of chemical compounds. The diversity analysis methods are used at the stage of compounds selection for screening and subsequent lead generation.

The application of these methods are based on the similar property principle,[1] according to which structurally similar molecules should manifest similar physicochemical and biological properties.

Structural similarity and nearest neighbors search along with substructure search has already been used for a long time.[2] Presently, it is a component of the majority of commercial chemical database management systems, such as the following: *MDL ISIS/Host*,[3] *Daylight Database Package*,[4] *CambridgeSoft ChemFinder*,[5] *Oxford Molecular's RS*[3] *Discovery*,[6] *Synopsys Accord*,[7] and *Tripos UNITY*.[8]

The detailed review of various aspects of similarity search in chemical databases is given in reference.[9] A review of commercial software designed for diversity analysis is presented in ref 10.

Many known approaches to the problem of similarity and diversity analysis differ by descriptors, similarity/diversity measures, and compounds selection algorithms.

For this purpose descriptors of various types[11] can be used, for example, topological indices, physical property descriptors, 2*D* and 3*D* structural fragments, etc.

The traditional whole-molecular descriptors, such as *logP*, pK_a, *molecular refractivity*, *dipole moment*, etc., are poorly appropriate for this task.

The 2*D* structural fragments are presently most widely used. As shown in many works,[12−14] the 2*D* fragments exceed, in most cases, descriptors of other types by their capability of recognizing biologically active and inactive compounds. Different types of atom-, bond-, and ring-centered two-dimensional structural fragments are known.[15] Structural keys use a predefined fragment dictionary. *CAS ONLINE Screen Dictionary*[16] and *MDL MACCS keys*[3] are examples of structural keys. Another types of *2D* fragments are automatically generated fragments, such as *Daylight*[4] and *Tripos UNITY*[8] *fingerprints*. In the fingerprinting algorithms all possible combinations of augmented atoms, atom sequences of specific length, and ring fragments are usually generated. Then, using pseudorandomizing algorithms, they are hashed into a fixed-length bit-string. Atom pairs,[17] topological torsion,[18] and autocorrelation indices[19] are also frequently used *2D* fragments descriptors.

Among *3D* descriptors, *CoMFA fields*[20] turned to be efficient[21] for similarity and diversity calculations. Other types of *2D* and *3D* descriptors are considered in detail in reviews.[9,11]

Tanimoto, cosine, and *dice similarity coefficients* and *Hamming* and *Euclidean distances* are most often used[9] as measures of similarity (or, by contrast, dissimilarity or distance measures) between two objects.

The majority of diversity measures are based on the above-mentioned pairwise intermolecular dissimilarity measures. *Minimum intermolecular dissimilarity, mean intermolecular dissimilarity*, and *average nearest neighbor distance* are examples of such distance-based diversity measures. *Minimum weight spanning trees*[22] also belongs to this class. The calculation time for all these functions at direct realization depends on the number of compounds according to a square-law. In particular case, when *mean intermolecular dissimilarity* and *cosine similarity coefficient* are used, a linear plot of diversity measure calculation speed vs number of compounds in the data set can be achieved using the centroid algorithm.[23] Alternative algorithms based on the *k-dimensional trees*[24] and *Kolmogorov-Smirnov statistic*[25] have recently been proposed. In the last case, the calculation time of the corresponding diversity coefficient is independent of the data set size.

* Corresponding author phone: (096)524-8269; e-mail: Sergey_Trepalin@chemdiv.com; trep@ism.ac.ru.
† Institute of Physiologically Active Compounds.
‡ ChemDiv, Inc.

10.1021/ci0100649 CCC: $22.00
Published on Web 02/01/2002

Dissimilarity-based compound selection algorithms[26] have become the most widespread for selection of structurally diverse subsets of compounds. First of all, these are different variants of *maximum dissimilarity selection algorithm.*[27] This is an iteration procedure at each step of which a compound maximally dissimilar to already selected compounds is selected. Depending on the definition of "maximum dissimilarity", different realizations of this algorithm are possible. For instance, in well-known *MaxMin* algorithm a compound with maximum nearest-neighbor distance to previously selected compounds is selected at every stage. And in the *MaxSum* algorithm, the compound is chosen in such a way that the sum of all distances of this compound to already selected compounds is maximum. In the general case, *maximum dissimilarity selection algorithm* has the time complexity $O(n^2N)$ for selecting a subset of the size n from a database of the size N. However, more efficient realizations of particular cases of this algorithm with a complexity of an order of $O(nN)$ for *MaxSum*[23] and *MaxMin*[28,29] are described.

Another type of dissimilarity-based compound selection algorithms is the *sphere exclusion algorithm*.[30] In this algorithm some similarity threshold is defined, which can be considered as a hypersphere radius in the multidimensional descriptors space. Then, the next compound is selected at every stage, and compounds that have similarity to this compound exceeding the specified threshold are excluded from further consideration. *Cluster sampling*[31] and *minimum dissimilarity selection*[32] exemplify this class of algorithms. The *optimizable K-dissimilarity selection*[32] procedure is described, which generalizes *maximum dissimilarity selection* and *sphere exclusion*.

In addition to the described above iteration procedures, the direct optimization of diversity objective functions is often used in combination with *stochastic algorithms.*[33] They include *simulated annealing,*[34] *genetic algorithms*,[35,36] and *evolutionary programming*.[37,38] The advantage of these methods is their versatility as they can be used with any quantitative diversity measures.

Statistical analysis and experimental design methods are also used for compounds selection. One of these methods is *D-optimal design.*[39] In this method compounds are selected so that potential errors in descriptors have minimal impact on supposed regression model. This is achieved by the maximization of the determinant D of the covariance matrix $X'X$ for the design matrix X, whose rows are compounds and columns are descriptors and/or their functions, e.g., squares. This is equivalent to the minimization of the inverse matrix determinant and the corresponding minimization of the prediction error of the assumed model. The use of *D*-optimal design results in the selection of compounds at the edges of the descriptor's space.[29]

Partition-based compound selection methods[40] have become widespread. They are based on the partition of the multidimensional chemistry space into hypercubic cells. Then compounds are assigned to one or several cells in this property space. Since the cell-based algorithms allow determining not only intermolecular distance but also the absolute position in chemistry space, they are perfectly suitable for finding and filling diversity voids as well as for other tasks related to diversity analysis. The calculation complexity of the partition-based algorithms does not exceed $O(N)$. However, practical realization of these algorithms is restricted by the chemistry space of a comparatively low dimensionality (usually not more than 10 dimensions). The *BCUT* values[41,42] turned out to be especially successful descriptors for this task.

The described above algorithms implemented in the commercial software[10] either do not solve or partially solve the following problems:

1. Diversity sorting should be performed taking into account compounds already available in the collection. This problem differs from classical diversity sorting because there already exists the original set of compounds which is not subject to sorting out and considered fixed.

2. Operation with large databases. A considerable computer time, sometimes several weeks, is needed for diversity sorting of the collection of 500 000 compounds by standard algorithms. As for databases with the size of more than million of compounds, this is a special type of problems, which requires expensive computer equipment and considerable time.

3. The available chemical compounds selection algorithms insufficiently take into account chemical uniqueness, which is significant for new drugs design. For example, when a compound contains a new, unique heterocycle and several standard substituents (phenyl, piperidine), it can be in the middle or even at the end of the sorted list.

DIVERSITY SORTING ALGORITHMS

The diversity sorting procedure is standard: initially structural fragments (screens) are generated for all compounds in the database. These screens are used to define the similarity measure $SIMILARITY(I,J)$ for any pair of compounds. The dissimilarity of a pair of chemical compounds is defined as $DISSIMILARITY(I,J) = 1 - SIMILARITY(I,J)$. As a diversity measure of the data set A with the size N, we use the average value of all pairwise intermolecular dissimilarities:

$$DIVERSITY(A) = \frac{\sum_{I=1}^{N}\sum_{J=1,I\neq J}^{N} DISSIMILARITY(I,J)}{N(N-1)} \quad (1)$$

Unlike most published algorithms, we use not linear but spherical fragments with different radii of the sphere around the chosen atom. The radius of the sphere is the topological distance between the central atom and the atom maximally remote from it, and each bond length is assumed to be equal to 1. To divide the molecule into spherical fragments, the center on each atom in the molecule is chosen in turn. Then, the fragments with the sphere radius equal to 1, 2, etc. are taken into account. The division into fragments of the 3-methylpiperidine molecule is shown in Figure 1.

It was empirically established that the sphere with radius 2 is sufficient to obtain a good result in the calculation of similarity/diversity coefficients. For the sphere with this radius, the average number of fragments generated from one molecule is equal to 34. This value was obtained from the database of 241 367 compounds, ChemDiv,[43] collection, which was developed taking into account diversity of its chemical compounds. For this database, the total number of different fragments (screens) is equal to 43 000. A less

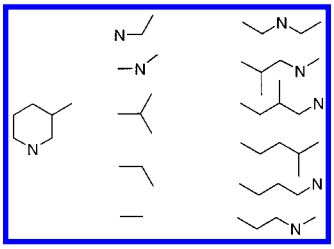


Figure 1. Fragments of structures obtained from 3-methylpiperidine for the sphere with radius 2.

number of screens is usually used in commercial programs. For example, *Daylight fingerprints*[4] by default have the size of at most 2058 bits. One can increase the default number of fingerprints, but the calculation time will consequently grow. *Tripos UNITY fingerprints*[8] have the length of 988 bits. *MDL MACCS keys*[3] use the library of 166 or 960 predetermined fragments.

The atom type (its position in the Periodical Table, charge, and valence) and bond type (single, double, triple, aromatic, coordination) are taken into consideration for each sphere. This type of division of a molecule into spherical fragments has previously been described.[44] The use of information on the cyclic kind of a bond is a specific feature of this algorithm. For example, the C—C fragment in the aliphatic chain differs from that in cyclopropane, which differs in turn from a similar fragment in cyclopentane. No difference is made for cycles with the size of 6 and higher: similar fragments in cycles with different sizes are considered to be equivalent. This method of division is also reasonable from the chemical point of view: it takes into account the unique behavior of small cycles in reactivity. In addition, aromatic bonds in five- and six-membered aromatic cycles are different.

Two algorithms are used for diversity sorting:

Maximum Dissimilarity Selection. 1. The choice of the first compound, which is most dissimilar all other compounds in the sorting list.

2. The choice of the compound, which is most dissimilar to already selected compounds. When we have several such compounds, the compound containing the maximum number of structural fragments is given the preference.

3. Point 2 is repeated until the list of all compounds is exhausted or the specified number of compounds is selected.

When this algorithm is used for the addition of compounds to the collection, Point 1 is omitted and the first compound is chosen as the most different from already available collection.

Maximum Similarity Elimination.[31] 1. Elimination of the most similar compound from the current set.

2. Procedure 1 is repeated until one compound remains.

When this algorithm is used for the addition of compounds to the collection, the elimination of compounds already present in the collection is forbidden. The end of calculation takes place when the remaining set is completely equivalent to the collection.

In both cases, a change in diversity of data set in the discrimination of each compound should be calculated

$$\Delta D = DIVERSITY(A_{N+1}) - DIVERSITY(A_N) \quad (2)$$

where $DIVERSITY(A_{N+1})$ is the diversity of the set of $N+1$ compounds, and $DIVERSITY(A_N)$ is the diversity of the set of N compounds. The efficiency of calculation of ΔD determines the speed of diversity sorting.

The chosen algorithm of ΔD calculation is described in ref 23. In this case, the angle cosine between the corresponding bitstring vectors M_I and M_J is used as a similarity measure of the pair of compounds I and J. The cosine, is a very acceptable similarity measure for bit vector, used elsewhere. However, for integer and real vectors special effort should be made to take the scale into consideration. The number of components F of these vectors is equal to the number of screens in the set, and the $M(K)$ component is equal to 1, if the Kth fragment is present in the compound and to 0 in the opposite case:

$$SIMILARITY(I,J) = \frac{\sum_{K=1}^{F} M_I(K)M_J(K)}{\sqrt{\sum_{K=1}^{F} M_I(K)^2 \sum_{K=1}^{F} M_J(K)^2}} \quad (3)$$

The eq 3 may be treated as the number of bits set in both molecules, normalized by the number of bits set in each molecule. In this algorithm the calculation of the sum of all pair similarities $SIMILARITY(I,J)$ between molecules in the set A containing N compounds is replaced by the dot product of two vectors A_C[23]

$$A_C \cdot A_C = \sum_{I=1}^{N}\sum_{J=1}^{N} SIMILARITY(I,J) \quad (4)$$

where A_C is the centroid vector, whose number of components is equal to the number of screens in the data set, and the numerical value of the Kth component is the following:

$$A_C(K) = \sum_{I=1}^{N} W(I)M_I(K) \quad (5)$$

where the weight of each Ith compound $W(I)$ is determined by the following formula:

$$W(I) = \frac{1}{\sqrt{\sum_{K=1}^{F} M_I(K)^2}} \quad (6)$$

The use of this algorithm considerably shortens the calculation time ΔD. When the traditional algorithm is used for the data set of N compounds, which is characterized by K fragments, the calculation time ΔD is proportional to NK. The calculation time in centroid algorithm is proportional to K and independent of N.

Our modifications of the algorithm[23] increase speed of calculation of ΔD. Consider ΔD as a function of the centroid vector

$$-\Delta D=(C_{set}+C_{n+1})(C_{set}+C_{n+1})/(N+1)^2-C_{set}C_{set}/N^2 \quad (7)$$

where C_{set} is the centroid vector for a data set of N compounds, and C_{n+1} is the centroid vector of $N+1$ compounds. In the algorithm *maximum dissimilarity selec-*

tion, we have to choose such an $N+1$ compound that ΔD is minimum. After simple mathematical transformations, we conclude that the dot product $C_{set}C_{n+1}$ should be *minimum*. Similar consideration of the *maximum similarity elimination* algorithm suggests that the compound, whose dot product $C_{set}C_{n+1}$ is *maximum*, should be excluded from the available data set of $N+1$ compounds.

Now let us examine the $C_{set}C_{n+1}$ product. The C vector has the K dimensionality: the total amount of screens in the database. To calculate the dot product, we have to perform K multiplication operations and (K-1) additions. For databases of several hundreds of thousands of compounds, K is of an order of 40 000, as mentioned above. However, in the C_{n+1} vector only the components corresponding to the screens, which are present in the $N+1$ compound, are nonzero. If the numbers of nonzero components are stored in the C vector for each compound, K_j multiplications and $(K_j\text{-}1)$ additions are required for the calculation of the dot product, where K_j is the number of screens in the j-th compound. As mentioned above, the K_j value is equal to 34 in average for ChemDiv database. Thus, taking into account the zero components of the centroid increases the speed of calculations by 1000 times. In addition, the storage of only nonzero components allows the efficient work with the computer memory. For the above example (database of 241 367 compounds; average number of screen: 34), 136 bytes of RAM per compound are required. For this method of information storage, large databases (1 000 000 compounds) can be loaded in 136Mb of RAM.

When the database contains M compounds and the current data set contains N compounds, to find the next $N+1$ data set element one should $(M-N)$ times calculate the $C_{set}C_{n+1}$ value. This assumed the square increase in the time required for diversity sorting of the database with its size. The performance of the algorithm can be improved by sorting of the remaining $(M-N)$ compounds by the $C_{set}C_{n+1}$ value. Indeed, if we fix a compound J and the data set is formed from the remaining compounds, the inequality is true: $C_{set(N)}C_j <= C_{set(N+1)}C_j$. When the data set size increases, the dot product of the centroid and vector of some compound either remains unchanged or increases. We can sort the remaining $(M-N)$ compounds by the $C_{set}C_j$ value. After this, the first compound in the sorted array is added to the data set, and this compound is removed from the sorted array. For the addition of the next compounds to the data set, new $C_{set}C_j$ values should be calculated for the L first sorted compounds. To find L, we calculate $C_{set}C_1$ (C_1 is the vector for the first compound in the sorted array) and, if its value differs from the previously calculated one, the bisections algorithm is used to find such an L index that $C_{set}C_L > C_{set}C_1$.

With the growth in the data set size, L increases and new $C_{set}C_j$ values should be calculated for the remaining compounds, and they should be repeatedly sorted by the value of this parameter. This is executed after the selection of P compounds. The calculation time as a function of the P parameter is presented in Table 1 for the database of 3300 compounds.

The optimum value of the P parameter depends on the $(M-N)$, the number of compounds, which must be sorted. We used the value $P = 256$ regardless of the $(M-N)$ value. At this parameter value, the calculation time increases by at most 30% compared to the optimum value of the P parameter for the data set with this size. This value of the parameter is optimum for the database of 100 000 records size.

Table 1. Calculation Time as a Function of the Sorting Repetition Frequency for the Data Set of 3300 Compounds

P	time (ms)	P	time (ms)
2	40865	64	27055
4	32680	100	28261
8	28526	150	29736
16	26735	200	31145
32	26161	infinite	43576
50	26706		

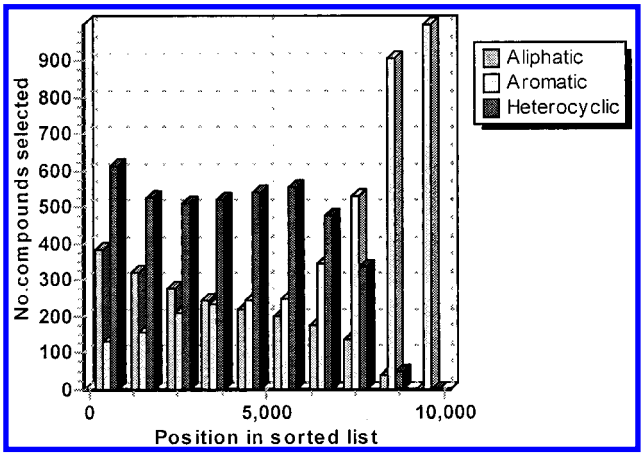


Figure 2. Histogram of the results of sorting of the testing data set by the *maximum dissimilarity selection* algorithm.

The addition of compounds to the available collection can be used to find possible errors in structures added. In fact, when the collection is rather representative and its chemical structures is thoroughly verified, the appearance of the new fragment with the topological radius 1 among the added structures indicates, in most cases, an error in the chemical structure: pentavalent carbon, etc. Therefore, when new compounds are added to the existing collection, flags are put out for all screens with the topological radius 1. If structures, whose fragments are not marked with flags, are present in the selected compounds, the operator is warned about a potential error.

To estimate the efficiency of diversity sorting, we used the data set of 2000 aliphatic compounds (A), 4000 aromatic compounds (B), and 4000 heterocyclic compounds (C). The results of sorting of the randomly mixed set by the *maximum dissimilarity selection* algorithm are presented in Figure 2.

Aromatic compounds, which are of minor interest, are efficiently eliminated from the data set: 2000 of them are placed at the end of the list, where aliphatic and heterocyclic compounds did not get. Since diverse aliphatic compounds were used for testing, heterocyclic and aliphatic compounds are not separated. However, heterocycles are more important from the point of view of searching for new classes of biologically active compounds. The method for the solution to this problem is described below.

To compare the results of diversity sorting by two described algorithms (*maximum dissimilarity selection* and *maximum similarity elimination*), we used the test database of 10 000 records in size. It contains 5000 *a fortiori* "interesting" chemical compounds. "Interesting" compounds were obtained by analyzing frequency of compounds sale in ChemDiv shop[43]—they attract at least 1.5 times more

Table 2. Comparison of the Distribution Histograms of "Interesting" Compounds for Diversity Sorting by the *Maximum Dissimilarity Selection* and *Maximum Similarity Elimination* Algorithms

	subset	1−1000	1001−2000	2001−3000	3001−4000	4001−5000	5001−6000	6001−7000	7001−8000	8001−9000	9001−10000
maximum dissimilarity selection	10000	740	753	698	591	533	513	363	335	242	232
maximum similarity elimination	10000	729	734	674	610	537	532	352	311	265	256

orders, than the average number of orders for a set. The other 5000 compounds were randomly chosen: they do not coincide with the collection with respect to which diversity sorting was performed. Diversity sorting of the testing set was carried out with respect to the available collection (241 367 records). The results of sorting were divided into 10 intervals 1000 compounds each. After sorting, it was found by each algorithm that 90% of compounds in each interval are identical (Table 2). Thus, both algorithms give approximately the same result. The *maximum dissimilarity selection* algorithm is preferential for the selection of up to 30% of compounds from the initial base (at the beginning of the sorting list), whereas the *maximum similarity elimination* algorithm works better for the selection of more than 70% of compounds (at the end of the sorting algorithm).

For each algorithm, approximately 2/3 of "interesting" compounds got into the first half of the sorted list. In the first half of the sorted list 1/3 of compounds from randomly selected data set is presented. Thus, using traditional diversity sorting algorithms, we increase the probability of "interesting" compound selection by 2 times.

SELECTION OF HETEROCYCLIC COMPOUNDS

Heterocyclic compounds are very attractive in search for new target structures. Successful binding with a receptor requires heteroatoms, which can form hydrogen or chemical bonds and their appropriate spatial arrangement for cooperative interaction. Conformations, with low probability of existence in acyclic compounds, are often stabilized in heterocycles. Therefore, any new heterocycles are promising candidates for tests.[45,46]

The published algorithms on chemical compounds diversity analysis do not take into account this specific feature of heterocycles. For diversity sorting by traditional algorithms, unique heterocycles, if they contain a great number of oftenmet substituents, can be at the end of the list and rejected for tests.

To solve this problem, compounds are divided into cyclic fragments containing heteroatoms. With this purpose, all acyclic bonds are primarily eliminated, and then single atoms and cycles containing no heteroatoms are removed. The remaining heterocyclic fragments are used as screens for diversity sorting by the algorithms described above (*maximum similarity elimination* and *maximum dissimilarity selection*). Compounds, containing no heterocycle, are eliminated from the data set.

The data set of 10 000 records consisting of N-heterocyclic bicyclic compounds was used for testing the described above algorithm. These compounds can conventionally be grouped as follows:

Compounds from the first group sufficiently well studied (and, as a consequence, presented in the ChemDiv collection)

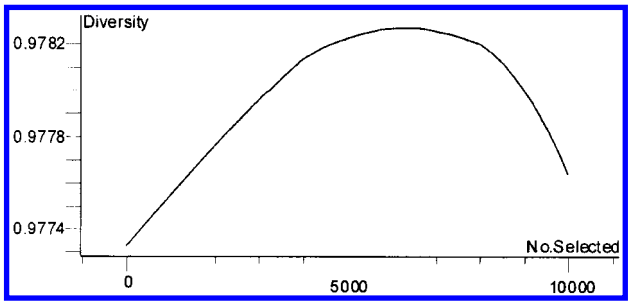
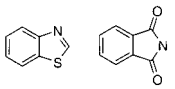


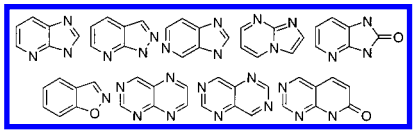
Figure 3. Diversity of the total "collection"+"test" data set vs number of selected compounds from "test."

many of which possess biological activity and are of interest from the pharmacological point of view:

2. Compounds widely presented in the collection but of minor interest:



3. New poorly studied compounds, which are few in the collection and promising for screening:



This testing data set ("test") was compared to the above-described collection ("collection") consisting of 241 367 records. The control data set ("control") was taken as controlling. It consisted of 10 000 compounds, and 5000 compounds are assigned to the first and third groups because they possess the known biological properties (drugs and their analogues) or are of interest due to their uniqueness.

Some compounds, along with the main N-bicyclic fragment, contained additionally one or several heterocycles (mainly thiophene, furan, piperidine, or pyridine). The change in the database diversity for choosing compounds from the "test" data set to "collection" vs the number of selected compounds is presented in Figure 3. The "test" data set was sorted by a standard procedure on the basis of analysis of the frequency of occurrence of heterocyclic fragments.

As can be seen in Figure 3, with the addition of new compounds from "test" to "collection," diversity of the total database first increases monotonically to some maximum and then decreases. This type of the curve is standard for sorting of a data set relative to the available collection.

Table 3. Distribution Histogram of Heterocyclic Compounds for Diversity Sorting of the Mixed Database Relative to the Collection[a]

heterocycle class	collection	test	1− 1000	1001− 2000	2001− 3000	3001− 4000	4001− 5000	5001− 6000	6001− 7000	7001− 8000	8001− 9000	9001− 10000
indoles	2998	733	346	120	5	9	380	14	23	34	47	23
benzimidazoles	5307	1268	58	192	344	213	126	79	269	29	84	109
benzothiophenes	4224	465	82	74	3	1	286	460	657	877	195	485
phthalimides	4818	374	91	220	344	635	62	11	65	22	94	58
quinolines	3465	595	152	89	1	0	146	427	10	14	98	64
1,4-dihydroquinolines	601	623	63	38	22	43	146	0	0	**26**	**309**	**239**
phthalazines	220	299	1	45	209	145	6	17	0	0	**22**	**3**
quinoxalines	566	568	110	163	72	0	5	1	1	**3**	**178**	**23**
purines	202	49	29	18	1	0	0	0	0	**2**	0	0
imidazo(4,5b)pyridines	24	41	2	42	0	0	0	0	1	**2**	**2**	0
pyrazo(3,4b)pyridines	38	27	28	0	0	0	0	0	0	0	0	0
imidazo(4,5c)pyridines	3	4	4	0	0	0	0	0	0	0	1	0
imidazo(1,2a)pyrimidines	51	27	2	7	0	2	1	**14**	**1**	0	**2**	**1**
2,3-dihydro-1h-imidazo(4,5b)-pyridin-2-one	12	4	4	0	0	0	0	0	0	0	0	**1**
benzisoxazoles	15	5	5	0	0	0	0	0	0	0	0	0
pteridines	20	10	8	0	0	0	0	0	0	0	0	**2**
pyrimido(5,4d)pyrimidines	1	1	1	0	0	0	0	0	0	0	0	0
dihydropyrido(2,3d)pyrimidin-7-one	0	17	10	4	0	0	0	0	0	**3**	0	0
total		5110	1011	1029	1013	1020	1023	1014	1013	1028	1017	1021
control		5000	750	653	539	487	493	438	431	401	385	360

[a] Substantially different values compared to those in Table 4 are emphasized by bold.

Table 4. Changes in the Distribution Histogram Taking into Account the Single, Rarest Heterocyclic Fragment

heterocycle class	collection	control	test	1− 1000	1001− 2000	2001− 3000	3001− 4000	4001− 5000	5001− 6000	6001− 7000	7001− 8000	8001− 9000	9001− 10000	1− 5000
1,4-dihydroquinolines	601	623	702	100	206	220	63	10	103	0	0	0	0	636
phthalazines	220	299	347	14	48	9	20	254	2	0	0	0	0	345
quinoxalines	566	568	701	50	61	83	156	120	151	68	0	17	0	470
purines	202	49	50	19	6	8	2	15	0	0	0	0	0	50
imidazo(4,5b)pyridines	24	41	49	4	0	12	33	0	0	0	0	0	0	49
imidazo(1,2a)pyrimidines	51	27	31	2	7	22	0	0	0	0	0	0	0	31
2,3-dihydro-1H-imidazo(4,5b)-pyridine-2-one	12	4	5	4	1	0	0	0	0	0	0	0	0	5
pteridines	20	10	10	2	0	8	0	0	0	0	0	0	0	10
7,8-dihydropyrido(2,3d)-pyrimidine 7-one	0	17	17	12	5	0	0	0	0	0	0	0	0	17

When we analyze the histogram of compounds selection by classes, the more additional fragments in the rows, the more complex the character of the plot. For example, for phthalimides, phthalazides, and quinoxalines containing virtually no other heterocycle, the selection curve is a parabola with one maximum, whereas for indoles, benzimidazoles, quinolines, and others, the selection curves are more complex (Table 3).

Thus, 1011 fragments were selected (column total) in the set of the first 1000 compounds, which is explained by the fact that some compounds contain two and more analyzed fragments.

It is unexpected that heterocycles poorly presented in the collection get to the end of the list. These data are emphasized by bold in Table 3. This is explained by the fact that these compounds contain, along with the main heterocycle, several often met heterocycles (furan, pyrrole, etc.). Therefore, heterocyclic compounds should be selected only taking into account the main, rarest heterocycle. The algorithm described above was modified as follows:

1. The temporary array for storing the occurrence frequency of screens of the *already selected* compounds is created.

2. For the calculation of ΔD when selecting a new compound, only one most rarely met fragment is used.

3. After the selection of each new compound, the occurrence frequency of *all* screens of this compound increases by 1.

This algorithm solved the described above problem. The results of calculation for some classes of compounds are presented in Table 4 as the interval distribution.

As can be seen in Table 4, all rare heterocyclic compounds are shifted to the beginning of the list, except for quinoxalines, which are widely presented in "collection," although being interesting from the chemical point of view.

The formalism allows a more correct selection of compounds. However, since *all* screens of a certain compound are used for diversity calculation rather than the single one, the diversity plot vs the number of selected compounds loses its monotonic character (Figure 4). To obtain a monotonic curve, diversity calculation should be performed under the assumption that each molecule contains one fragment.

Organic compounds can be treated as a number of the central heterocyclic (or aromatic, cyclic) fragment, aliphatic chain linking cyclic fragments between each other, and functional groups. The properties of the heterocycle is substantially affected by its nearest environment and, first of all, the presence of double bonds connected with the heterocycle. In most cases, this combination of the heterocycle with the keto (thio or imino) group should be considered as an independent screen as a whole (for example,

Table 5. Comparison of the Efficiency of Ketopiperazines Sorting Taking into Account and Ignoring the Double Bond for Screen Generation[a]

	collection	"piper"	1−1000	1001−2000	2001−3000	3001−4000	4001−5000	5001−6000	6001−7000	7001−8000	8001−9000	9001−10000
Double Bond Is Taken into Account												
no. of compds without keto groups	8691	9750	873	923	954	1000	1000	1000	1000	1000	1000	1000
no. of compds with keto group	100	250	**127**	**77**	**46**	0	0	0	0	0	0	0
Double Bond Is Not Taken into Account												
no. of compds without keto groups	8691	9750	960	948	971	986	992	982	983	982	987	959
no. of compds with keto group	100	250	40	52	29	14	8	**18**	**17**	**18**	**13**	**41**

[a] Substantially different values are printed in bold.

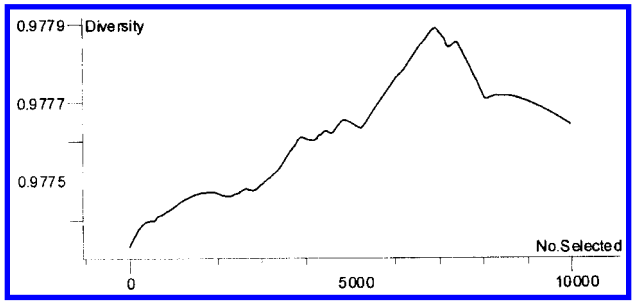


Figure 4. Diversity as a function of the number of selected compounds for selection by the single, rarest heterocycle.

barbituric acid, purines, coumarins, flavones, and many others). Therefore, atoms linked to the cycle by the double bond can be left as an option for heterocyclic screens generation. The testing calculations were performed for the model data set ("piper") containing 10 000 piperazines, ketopiperazines, and 2,3-, 2,5-, and 2,6-diketopiperazines. The collection relative to which the diversity sorting was performed contains few keto- and diketopiperazines. The results of calculations are presented in Table 5.

When the connected double bond at the heterocycle is ignored, ketopiperazines are randomly selected along with piperazines. When the double bond is taken into account, all of them are shifted to the beginning of the list.

Many heterocycles represent bi-, tri-, and polycyclic systems formed by heterocycle condensation with benzene, cyclohexane, or other cycle containing no heteroatom. From the chemical point of view, these compounds often resemble the initial systems. This is especially pronounced for systems of the type thiophene-tetrahydrobenzothiophene or 2,5-dihydro-1H-2,5-pyrroledione-phthalimide, where the addition of cyclohexane or unsubstituted benzene ring does not substantially change the properties of the condensed compound. In several cases (especially for condensation with substituted benzene), the distinctions can be more substantial. This requires the introduction into the algorithm of an option, which allows one to take into account or ignore condensed cycles containing no heteroatom.

Analyzing heterocycles containing condensed cycles, we sometimes have to restrict the number of cycles, i.e., to consider not the whole condensed system with the heteroatom but isolate the basic heteroatom-containing cycle. This is reasoned by the uniqueness of virtually any condensed cycle with the number of cycles more than 5, and these compounds are always selected into the collection. At the same time, their selection is not desirable because most similar compounds are characterized by high lipophilicity coefficients *LogP* and molecular weight and, as a consequence, low solubility and membrane permeability. Thus, they are bad candidates for biological screening. According to the Lipinskii rule,[47] for compounds with more than five condensed rings, $MW > 500$ and $LogP > 5$, biologically active compounds are at most 20%. Depending on the number of compounds containing condensed cycles, the optimum maximum of the number of accounted cycles in the analyzed data set can be different. It was empirically found for the testing data set of 10 000 condensed phthalimides that condensed cycles with the number of cycles (N_{max}) of at most 3 should be taken into account. For the data set of 10 000 compounds, various imidazole derivatives, the optimum value is $N_{max} = 4$. Therefore, the maximum number of cycles in polycyclic compounds was introduced as an option.

COMPOUNDS SELECTION FOR TESTS

In many cases, it was of interest to obtain the database of chemical compounds resembling as much as possible the data set consisting of one or several specified structures (chemical analogues), for example, biologically active. To solve the problem, the data set of compounds with given activity is divided into clusters. If there exists a number of distinct clusters, the initial data set is divided into subsets *L*. They contain compounds, belonging to single cluster. Otherwise single data set *L* is formed for each type of biological activity.

Then similarities of each compound selected to the collection and each compound in this data set are calculated and the average similarity is used

$$s_{i,set} = \Sigma s_{ij}/N \tag{8}$$

where $s_{i,set}$ is the average similarity of the *i*th compound in the collection with the data set *L*, s_{ij} is the similarity of the *i*th compound in the collection with the *j*th compound in *L*, and *N* is the number of compounds in this set.

The time of $s_{i,set}$ calculation by the classical scheme depends linearly on the data set size *N*. However, using cosines as a similarity measure and the centroid algorithm, we can write the equation for $s_{i,set}$ in the form

$$s_{i,set} = C_{set}C_i/N \tag{9}$$

The expensive calculation of coefficients s_{ij} is replaced by the dot product of two centroid vectors: for the data set with the specified biological activity C_{set} and for the *i*-th compound from the collection C_i. Taking into account that C_i is a rare vector (on the average, 34 components), 34 multiplications and 33 additions are needed for $s_{i,set}$ calculation. It is independent of both the data set *L* size and the total amount of screens in the base. The calculation time of the C_{set} vector depends linearly on the size *L* but it is calculated only once. If the collection size is *M*, the total

Table 6. Comparison of Results of Similar Compounds Selection Using Daylight Software[4] and Algorithms (ChemoSoft) Described in This Work

	1−1000	1001−2000	2001−3000	3001−4000	4001−5000	5001−6000	7000−241367
Chemo-Soft	852	59	40	35	11	3	0
Daylight	865	72	31	27	5	0	0

time of calculation by the centroid algorithm is proportional to $M+N$, whereas calculation by the classical method required time expenses of an order of MN.

The obtained $s_{i,set}$ coefficients are used for candidate selection for testing. The absolute value of the similarity $s_{i,set}$ is bad for the selection of promising candidates. The data set L can be nonuniform, if several classes of compounds exhibit this biological activity. To determine whether the compounds from collection L is promising for testing by this biological activity or not, the average similarity coefficient of compounds in the data set L is calculated

$$s_L = \Sigma s_{ij}/(N(N-1)) \quad (10)$$

Summation is executed over all unequal indices i,j. The same equation written through the centroid vector is the following:

$$s_L = C_{set}C_{set}/(N(N-1)) \quad (11)$$

Compound i is considered promising for testing by this type of biological activity if $s_{i,set} >= s_{L.}$

It was empirically found that the efficient selection of homologues can be performed at the diversity coefficient of the database lower than 0.5 (the similarity coefficient should be higher than 0.5). For algorithm testing, we developed the data set "coumarin" of 10 000 derivatives of 8-hydroxycoumarins, and the collection of 241 367 compounds was similarity sorted relative to this data set. Structure fragment search showed that this collection contains 1000 derivatives of 7-hydrocoumarins and its bioisosteric analogues (similarity coefficient 0.54). The results of similarity sorting of the collection relative to the data set "coumarin" are presented in Table 6.

For comparison, Table 6 contains the results obtained by processing of the clusters generated by Daylight Clustering Package.[4] The results indicate a satisfactory closeness of both methods with respect to the real number of compounds with this scaffold in the collection. The calculation time using the ChemoSoft program is shorter than 10 min.

ALGORITHMS REALIZATION

The above-described algorithms are implemented in the ChemoSoft program, which is the further development of the CheD[48] program. ChemoSoft is a file-server of chemical databases containing initial information for calculations. Results of calculations are also stored in ChemoSoft-supported databases.

Diversity sorting and similarity calculation are formed as the intuitive drag-and-drop manner. If another database is the target, it is considered fixed and diversity sorting is performed relative to this database. The most diverse compounds are selected into the available collection by the same method. If the same database is the source and target, traditional sorting is performed with the purpose to obtain a representative data set.

Before diversity sorting, it is desirable to perform the primary search and remove identical chemical structures. If duplicates contain rare fragments, after sorting they may occur at the beginning of the sorted list. Duplicates elimination is provided as an option.

Similarity and diversity calculations are performed both for the entire contents of the database and for data sets as well. This is especially efficient for compounds selection for tests. One database contains compounds with different biological activities. A data set for a specified biological activity is formed from this database, and target compounds are sorted by similarity relative to this data set.

A separate modulus in ChemoSoft is used for compounds selection by new heterocyclic fragments. Diversity sorting of heterocycles can be performed for both a single base (to obtain a representative data set) and a reference database (compounds selection from external sources). Option parameters are screens generation with heteroatoms connected by the double bond; sorting by single screen, which is most rarely met at current stage of sorting; elimination of cycles containing no heteroatom; and elimination of polycyclic heterocycles provide a convenient tool for the solution to various problems.

Calculation results are presented in the following form:

1. For a comparison of two databases by similarity, the database is sorted in order of similarity decreasing. The numerical value of similarity is assigned to each recording. The sorted database or its part can be stored as a list together with numerical values of similarity coefficients.

2. For diversity sorting, the database is sorted in order of diversity decreasing. As for similarity sorting, data can be stored as a list.

3. Results of diversity sorting are derived in the form of the diversity as a function of the number of selected compounds. This plot can be stored and then retrieved after new sorting, which enables a visual comparison of the sorting results of several databases. For diversity sorting relative to the collection, records containing new fragments compared to the collection are indicated in the plot (points in Figure 5). These records can contain errors in the chemical structure. The tools for their browsing and removal of erroneous records are available (Figure 5).

The plot of the diversity vs the number of selected compounds descends monotonically for sorting of a single database. The following types of plots are possible for sorting relative to the collection:

1. The curve ascends monotonically. These can be very good data, which should be placed into the collection as a whole. However, more often this indicates a poor formation of the collection, ignoring structures diversity. For example, if the collection contains only aliphatic compounds, the addition of any aromatic compound, including duplicates of already added structures, results in diversity increasing. In this case, sorting relative to the whole collection should not be performed, it is necessary to create a representative data set and perform sorting relative to it.

2. The curve descends monotonically. The database contains classes of compounds often met in the collection. These data are rejected if the collection is formed taking into account diversity.

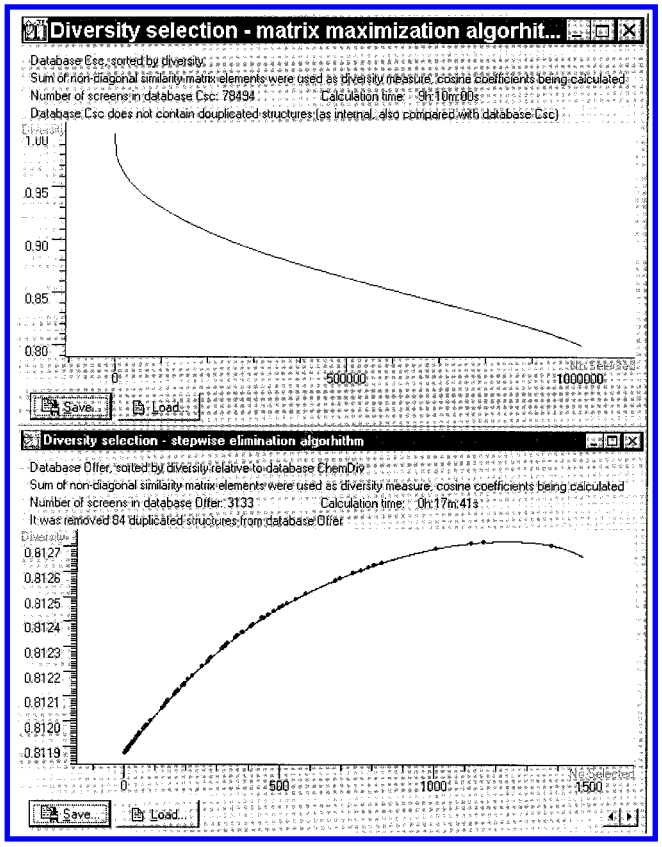


Figure 5. Representation of results of diversity sorting in the ChemoSoft program.

3. The curve primarily ascends and then descends. This is most common type of plots. Compounds corresponding to the maximum in the curve are selected for the formation of the maximal diverse collection.

CONCLUSION

The work of the program requires the Windows95/98/NT/2000 operation system. 256Mb RAM is enough for processing databases of more than 1 000 000 compounds without memory swapping. The time necessary for calculation is very short for this kind of problems. To choose new data from the initial base of 4000 records to the collection consisting of 213 000 records, 3 min 18 s are required for the *maximum similarity elimination* algorithm and 4 min 31 s for *maximum dissimilarity selection* (Pentium-III, 600 MHz). The time includes the search and removal from the initial database of identical structures (700 duplicates for the above example) and data reading from the disk. The time depends slightly on the collection sizes: approximately 1 min is additionally required for each 100 000 records in the collection.

At most, 9 h of the computer time (Pentium III, 800 MHz) are required for complete diversity sorting of the database of 1 008 343 records (ChemDiv, Inc. available database of real compounds synthesized in more than 200 institutes of the Academy of Sciences of the former USSR).

The efficiency of the algorithms is confirmed by the comparison of the calculation time with published data. For example, to select 461 most diverse compounds from the database of 90 000 compounds, 2 min 24 s are required (Pentium II, 500 MHz) using the described above *maximum similarity elimination* algorithm. To select 400 compounds of the 90 000 base using the simulated annealing algorithm and Kolmogorov−Smirnov statistics, 8 min are required (2 processors Pentium III, 800 MHz).[25] The time necessary for calculation by the same algorithm (simulated annealing) and mean nearest-neighbor distance[24] was estimated as 73 min.[25]

ACKNOWLEDGMENT

The authors thank A. V. Yarkov and A. V. Skorenko for help and useful discussions

REFERENCES AND NOTES

(1) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990.
(2) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nearest Neighbor Searching in an Online Chemical Structure Search System. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36−41.
(3) MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577, USA, http://www.mdli.com/.
(4) Daylight Chemical Information Systems, Inc., 441 Greg Avenue, Santa Fe, NM 87501, USA, http://www.daylight.com/.
(5) CambridgeSoft Corporation, 100 CambridgePark Drive, Cambridge, MA 02140, USA, http://www.camsoft.com/.
(6) Oxford Molecular Ltd., Medawar Centre, Oxford Science Park, Sandford-on-Thames, Oxford, OX4 4GA, UK, http://www.oxmol-.co.uk/.
(7) Synopsys Scientific Systems Ltd., 175 Woodhouse Lane, Leeds, LS2 3AR, UK, http://www.synopsys.co.uk/.
(8) Tripos, Inc., 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA, http://www.tripos.com/.
(9) Willett, P.; Barnard J. M.; Downs G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(10) Warr, W. A. *Commercial Software for Diversity Analysis. Perspect. Drug Discovery Des.* **1997**, *7/8*, 115−130.
(11) Brown, R. D. *Descriptors for Diversity Analysis. Perspect. Drug Discovery Des.* **1997**, *7/8*, 31−49.
(12) 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.
(13) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(14) Matter, H. Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of 2D and 3D Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.
(15) Adamson, G. W.; Cowell, J.; Lynch, M. F.; McLure, A. H. W.; Town, W. G.; Yapp, A. M. Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files. *J. Chem. Doc.* **1973**, *13*, 153−157.
(16) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. The CAS Online Search System. 1. General System Design and Selection, Generation, and Use of Search Screens. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93−102.
(17) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(18) Nilakantan, R.; Bauman, N.; Dixon. J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.
(19) Moreau, G.; Broto, P. The Autocorrelation of a Topological Structure: A New Molecular Descriptor. *Nouv. J. Chim.* **1980**, *4*, 359−360.
(20) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(21) Cramer, R. D.; Clark, R. D.; Patterson, D. E.; Ferguson, A. M. Bioisosterism as a Molecular Diversity Descriptor: Steric Fields of Single "Topomeric" Conformers. *J. Med. Chem.* **1996**, *39*, 3060−3069.
(22) Mount, J.; Ruppert, J.; Welch, W.; Jain, A. N. IcePick: A Flexible Surface-Based System for Molecular Diversity. *J. Med. Chem.* **1999**, *42*, 60−66.
(23) Holliday, J. D.; Ranade, S. S.; Willett, P. A Fast Algorithm for Selecting Sets of Dissimilar Molecules from Large Chemical Databases. *Quant. Struct.-Act. Relat.* **1995**, *14*, 501−506.

(24) Agrafiotis, D. K.; Lobanov, V. S. An Efficient Implementation of Distance-Based Diversity Measures Based on k−d Trees. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 51−58.
(25) Agrafiotis, D. K. A Constant Time Algorithm for Estimating the Diversity of Large Chemical Libraries. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 159−167.
(26) Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of Algorithms for Dissimilarity-Based Compound Selection. *J. Mol. Graph. Mod.* **1997**, *15*, 372−385.
(27) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137−148.
(28) Polinsky, A.; Feinstein, R. D.; Shi, S.; Kuki, A. *LiBrain: Software for Automated Design of Exploratory and Targeted Combinatorial Libraries. In Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery*; Chaiken, I. M., Janda, K. D., Eds.; ACS Conference Proceeding Series, 1996; pp 219−232.
(29) Higgs, R. E.; Bemis, K. G.; Watson, I. A.; Wikel, J. H. Experimental Designs for Selecting Molecules from Large Chemical Databases. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 861−870.
(30) Hudson, B. D.; Hyde, R. M.; Rahr, E.; Wood, J. Parameter Based Methods for Compound Selection from Chemical Databases. *Quant. Struct.-Act. Relat.* **1996**, *15*, 285−289.
(31) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59−67.
(32) Clark, R. D. OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181−1188.
(33) Agrafiotis, D. K. Stochastic Algorithms for Maximizing Molecular Diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 841−851.
(34) Zheng, W.; Cho, S. J.; Waller, C. L.; Tropsha, A. Rational Combinatorial Library Design. 3. Simulated Annealing Guided Evaluation (SAGE) of Molecular Diversity: A Novel Computational Tool for Universal Library Design and Database Mining. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 738−746.
(35) Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: Reading, MA, 1989.
(36) Gillet, V. J.; Willett, P.; Bradshaw, J. The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731−740.
(37) Clark, D. E.; Westhead, D. R. Evolutionary Algorithms in Computer-Aided Molecular Design. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 337−358.
(38) So, S. S.; Karplus, M. Evolutionary Optimization in Quantitative Structure−Activity Relationship: An Application of Genetic Neural Networks. *J. Med. Chem.* **1996**, *39*, 1521−1530.
(39) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.
(40) Mason, J. S.; Pickett, S. D. Partition-Based Selection. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 85−114.
(41) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspect. Drug Discovery Des.* **1998**, *9*, 339−353.
(42) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(43) ChemDiv, Inc. 11575 Sorrento Valley Road, #210, San Diego, CA 92121, USA, http://www.chemdiv.com/.
(44) Bremsler, W. HOSE-A Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355−365.
(45) Nicolaou K. C., Pfefferkorn J. A., Barluenga S., Mitchell H. J., Roecker A. J., Cao G.-Q. Natural Product-like Combinatorial Libraries Based on Privileged Structures. 3. The "Libraries from Libraries" Principle for Diversity Enhancement of Benzopyran Libraries. *J. Am. Chem. Soc.* **2000**, *122*, 9968−9976.
(46) Abrous L. I., Hynes J., Jr., Friedrich S. R., Smith A. B., Hirschmann R. Design and Synthesis of Novel Scaffolds for Drug Discovery: Hybrids of â-D-Glucose with 1,2,3,4-Tetrahydrobenzo[*e*][1,4]diazepin-5-one, the Corresponding 1-Oxazepine, and 2- and 4-Pyridyldiazepines. *Org. Lett.* **2001**, *3*, 1089−1092.
(47) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug. Delivery Rev.* **1997**, *23*, 3−25.
(48) Trepalin, S. V.; Yarkov, A. V. CheD: Chemical Database Compilation Tool, Internet Server, and Client for SQL Servers. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 100−107.

CI0100649