DISE: Directed Sphere Exclusion

Alberto Gobbi and Man-Ling Lee*

Anadys Pharmaceuticals Inc., San Diego, California 92121

Received June 18, 2002

The Sphere Exclusion algorithm is a well-known algorithm used to select diverse subsets from chemical-compound libraries or collections. It can be applied with any given distance measure between two structures. It is popular because of the intuitive geometrical interpretation of the method and its good performance on large data sets. This paper describes Directed Sphere Exclusion (DISE), a modification of the Sphere Exclusion algorithm, which retains all positive properties of the Sphere Exclusion algorithm but generates a more even distribution of the selected compounds in the chemical space. In addition, the computational requirement is significantly reduced, thus it can be applied to very large data sets.

INTRODUCTION

Developing new drugs, in general, starts with screening hundreds of thousands of molecules against selected biological targets. Subsequent hits become lead molecules and finally, after many modifications, drug candidates and drugs. This process is very cost intensive. Therefore, it is important to acquire as much confirmed structure−activity information as possible from the primary screening in the early stages of the drug discovery process.

Major pharmaceutical companies have compound libraries with more than one million compounds for screening. Most of the compounds have been made for specific projects and are therefore highly clustered. Subsets of these collections need to be compiled for primary screening, to avoid redundancy. Biotech companies, on the other hand, must acquire a collection of compounds for primary screening from the huge number of compounds available from commercial vendors. Both pharmaceutical and biotech companies have to rely on appropriate compound selection algorithms to compile a suitable subset for primary screening.[1,2]

Selecting compounds for subsets is always a balance between diversity and coverage. The goal is to select the compounds such that confirmed lead structures and eventually some preliminary Structure Activity Relation (SAR) information[3] can be obtained from the hits and at the same time to avoid redundancy. To achieve this goal the selected compounds should be distributed as evenly as possible in the given chemical space, so that each compound has the necessary number of neighbors to provide confirmed information of interest.

Numerous selection algorithms have been published for various purposes. They, in principle, can be assigned to one of the three basic techniques: clustering, grid based selection, and distance based selection.

Clustering algorithms group the candidate compounds into clusters based on a distance measure.[4,5] Subsequently, one or more compounds are selected from each cluster to represent the other molecules in the cluster as well as their properties. Outliers, which cannot be assigned to any cluster, typically pose a problem when a clustering method is used for compound selection. The performance of most clustering algorithms is quadratic or higher in the number of candidate compounds since the distance of each compound to every other compound is measured to decide if it should be placed into the same cluster. Clusters typically vary in shape, size and density, making it difficult to control the regularity of the distances between the selected compounds.

Grid based selection[6,7] begins by partitioning a n-dimensional space covered by the candidate molecules into a grid of (hyper) cubes. As with clustering, one or more compounds are then picked from each cube to represent all compounds in that cube. One advantage of this class of methods is that sparsely- or nonpopulated regions in chemical space can be easily identified. Grid based selection requires the definition of a coordinate system for the properties of each molecule. The number of cubes rises as the power of the dimensionality. To keep the number of the cubes manageable, some properties are often combined with each other using techniques such as principle component analysis (PCA). The BCUT descriptors introduced by Pearlman[8] based on work by Burden[9] have been shown to be a good choice for this, because they capture a lot of information about a molecule.[7] Interpreting the meaning of the BCUT descriptors is however not straightforward. Grid-based algorithms are usually linear in the number of candidate compounds and therefore very fast. Given the coordinates of a compound, the cube to which it belongs can be readily computed. An even coverage of chemical space is easily achieved if it is possible to select compounds at or near the center of each cube.

Distance-based selection algorithms[10] rely only on a distance (similarity) measure between pairs of candidate compounds. This distance measure is used to select compounds, which are not near (similar) to each other. Distance-based algorithms include minmax,[11] D-optimal design,[12] and variations of the Sphere Exclusion algorithm.[1,13−16] Distance-based algorithms are easy to understand and to explain in terms of the distance defined between the compounds. The performance requirements are usually between those for clustering and grid-based methods. If N is the number of

* Corresponding author phone: +1 (858) 530 3658; e-mail: mlee@anadyspharma.com.

10.1021/ci025554v CCC: $25.00
Published on Web 11/28/2002

candidates and n the final number of selected compounds, the CPU requirement of the minmax algorithm is about O(n x N), since each selected compound needs to be compared to all candidates. Popular variants of the Sphere Exclusion methods, OptiSim[16] and DIVPICK,[15] reduce the CPU requirements by introducing a faster random selection step. The regularity of coverage of the selected subsets varies from one method to another.

In the past, emphasis in compound selection was on compiling diverse subsets. However, we believe that in the case of compound selections for primary screening, focus should be set on even coverage of the chemical space. Even coverage will ensure that hits found in the screens are not isolated structures, but instead representatives of analogue sets of compounds, which give information for lead structure identification and preliminary SAR analysis.[3]

In this paper, we describe a new variation of the Sphere Exclusion algorithm, Directed Sphere Exclusion (DISE), which achieves a more evenly distributed compound subset by abandoning the random selection approach and instead imposing a directed selection. It still shares the very easy to understand method with the other algorithms based on Sphere Exclusion, which iteratively excludes any compound within a given radius from any already selected one. In addition, DISE reduces CPU requirements significantly making it suitable for selection of compounds out of very large data sets.

METHOD

1. Sphere Exclusion. Since the DISE method was derived from the Sphere Exclusion algorithm, a brief description of this method is given first. Sphere Exclusion is a well-known dissimilarity-based compound selection method first described by Hudson et al.[13] and then later optimized by various groups.[14−16] The idea behind the algorithm is to select molecules, whose similarities with each of the other selected molecules are not higher than the defined threshold. Therefore, each selected molecule creates a (hyper) sphere around itself, so that any candidate molecules inside the sphere are excluded from the selection. The radius of the sphere is an adjustable parameter, determining the number of compounds selected and the diversity among them. The original method starts with the "most descriptive compound" and in each cycle identifies the compound most similar to the centroid of the already selected compounds. This requires a resorting of the candidate compounds after each selection and is very CPU intensive. Variations from the original algorithm, e.g. DIVPICK,[15] OptiSim[16] have been implemented. They reduce the CPU time required by selecting the next compound quicker and have become more widespread than the original method. Because of this, we have implemented the DIVPICK algorithm for comparison purpose and referred to this implementation as Sphere Exclusion (SE).

Shown in Figure 1 is the flowchart for the Sphere Exclusion algorithm.[1] Having defined the similarity threshold S_{max} (1 − exclusion sphere radius), the first molecule in the FILE with candidate molecules is transferred to the SUBFILE. All subsequent candidate molecules in FILE are compared against the selected molecules in SUBFILE. As soon as a selected molecule in SUBFILE has a similarity larger than S_{max}, the candidate molecule is rejected because

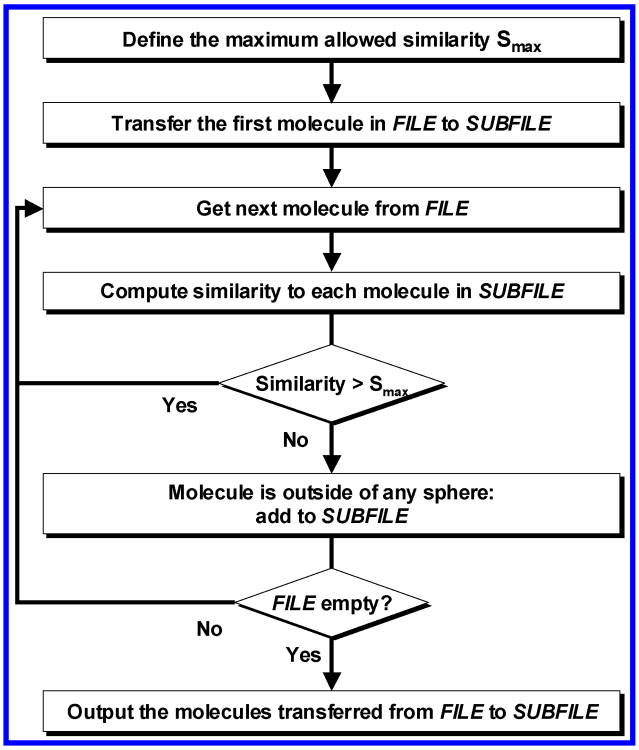

Figure 1. Sphere Exclusion Algorithm.

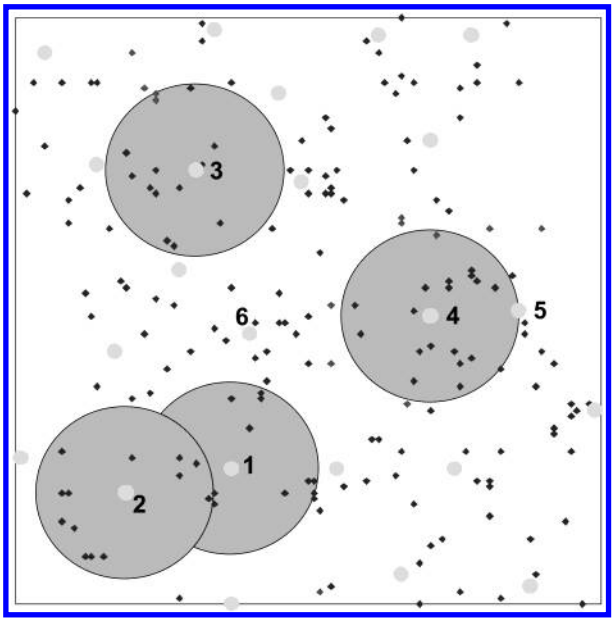

Figure 2. Sphere Exclusion − Simulated Selection. Selected points are given as (●). The first four spheres are shown as well as the selection order of the first six points (numbers).

it is inside this sphere. If it is outside the spheres of all molecules in SUBFILE, the candidate molecule is added to SUBFILE. The evaluation loop is repeated until no molecule is left in FILE.

Figure 2 shows the result of a Sphere Exclusion simulation. A set of 100 points with random 2D coordinates was generated and the Cartesian distance was used to run a Sphere Exclusion selection. As can be seen the distances between the selected points varies. The distance between the selected points **4** and **5** is the smallest possible one, **5** being just on the border of sphere **4**. Other pairs are sometimes farther apart from each other, as in the case of **1** and **6** or **4** and **6**.

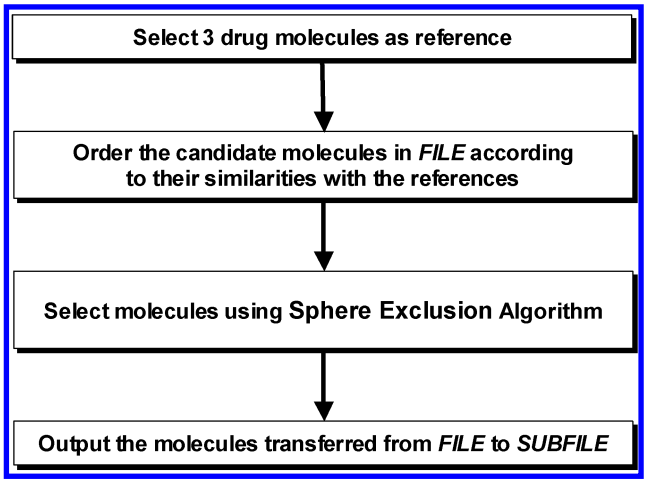

Figure 3. DISE − Algorithm.

This is because the selection is driven by the random ordering of points in the input set.

Recalling that the goal is to compile a set of compounds that evenly covers the area of chemical space, the ideal case would be if the distance between each molecule and its neighbors were always equal to the maximum allowed similarity S_{max}. DISE (DIrected Sphere Exclusion), a variation of the Sphere Exclusion algorithm, drives the selection toward this ideal case.

DISE abandons the idea of randomly picking the next compound from the candidate FILE for evaluation. Instead, prior to the actual selection with the sphere exclusion algorithm, the compounds in the candidate FILE are reordered according to their descending similarities with three reference molecules. The first step consists of sorting by similarity to the first reference (R1). This results in ties[17] whenever two candidate molecules have the same similarity to R1. The ties are resolved by sorting the tied compounds by descending similarity to R2 and finally by descending similarity to R3 (Figure 3). This yields a pre-clustering of the candidate molecules, which places molecules with similar structures next to each other in the input FILE. Since the function of the reference molecules is only to introduce a pre-defined order, the choice of the reference molecules is not very important. However, the reference molecules should be sufficiently dissimilar to each other.

Figure 4 shows a simulated DISE selection of points from the same set as in Figure 2. The algorithm steadily selects points with decreasing similarity to the first reference point R1. R2 and R3 are only important to break ties.[17] If there are points with identical distances to R1, the point nearest to R2 will be selected first. As can be seen by the numbering in Figure 4, the selection occurs in spheres around R1. Because of the pre-sorting, the selected points are very close to the border of the spheres of the already selected compounds. This yields an even coverage of the space provided there are enough points from which to choose. By selecting the points more evenly, the number of compounds in the subset compiled by DISE is larger than the number of compounds in the subset selected by the Sphere Exclusion method. In the simulations (Figures 3 and 4), 28 points were selected with the DISE algorithm while the Sphere Exclusion only selected 23. The method suggested by Hudson et al.[13] would result in a similar selection but is much more demanding in CPU time.

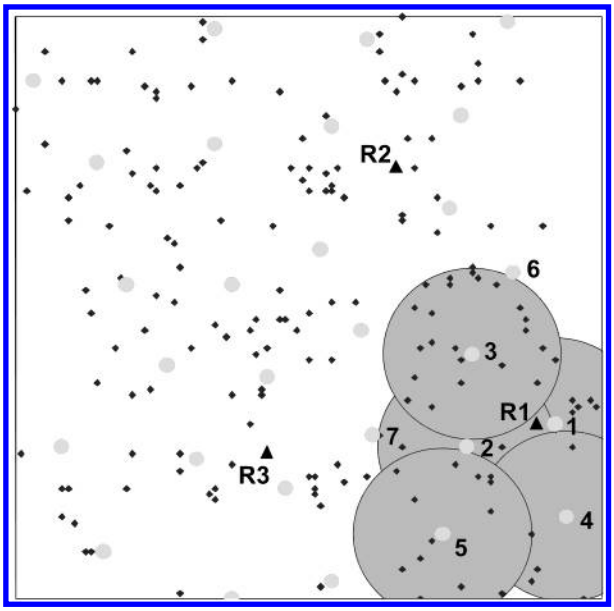

Figure 4. DISE − Simulated Selection. The three reference points (▲: R1, R2, R3) were added for display, but not as part of the candidate set. Selected points are given as (●). The first five spheres are shown as well as the selection order of the first seven points.

When implementing the DISE method, one additional benefit can be drawn from the pre-sorting of the input file: since compounds similar to each other will be near each other in the input file, rejection of compounds, which are inside the spheres, can be done earlier. To achieve this, comparison of a new candidate molecule starts with the last selected molecule from SUBFILE. This yields a significant performance improvement.

DETAILS

The Sphere Exclusion and DISE methods were implemented as a combination of Perl scripts and C programs. The Daylight toolkit[18] was used to compute 512 bit long fingerprints and to compute similarities based on the Tanimoto coefficient of these fingerprints.

For all selections, a threshold similarity of 0.85 corresponding to a sphere radius of 0.15 was used based on suggestions by Brown et al.[19] and Patterson et al.[20]

All CPU times given are for runs using a single CPU on a LINUX workstation running at 860 MHz.

DATA

In this study four data sets where used. The first one consists of the collection of commercially available compound from ASINEX.[21] Two preprocessing steps were applied to the original data set: 1) Compounds with molecular weight above 700 g/mol were removed. 2) A list of substructure queries compiled in-house was matched to the remaining compounds in order to identify and remove reactive and toxic fragments. Further filtering, e.g. using the "rule of 5", was not done per agreement with medicinal chemists, since we believe that physical chemical properties should not be considered in the selection of compounds for the primary screening set but in later stages of lead optimization. This is a typical set of compounds offered to

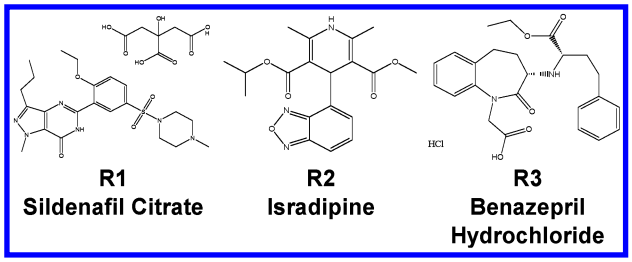

Figure 5. Reference compounds used for the DISE algorithm.

pharmaceutical companies for screening and will be referred to as data set A.

Some of the compounds in the first data set turned out to be singletons, meaning that none of the other compounds in the data set had a higher similarity than 85%. Both the DISE and Sphere Exclusion methods select all of them, because they do not fall in the sphere of any other compound. Some of these compounds have exotic structures, which are not interesting for screening, others however just happen to be the only compound out of a given class in the data set. To exclude the effect of the singletons on the results, two additional data sets were compiled: Set B includes all compounds (155 548) from set A which have at least one neighbor with a tanimoto similarity of 0.85 or greater. Set C includes all compounds (64 271) from set A, which have at least 20 neighbors with a Tanimoto similarity of 0.85 or greater. Set C is restricted to more densely populated areas of the chemical space filled by the ASINEX compounds. It is not suggested that these preprocessing steps, which are expensive in terms of computing time, are applied to any real selection. Sets B and C are used here only to highlight the improvements of the DISE algorithm over the Sphere Exclusion algorithm.

To evaluate the CPU requirements and behavior of the DISE method on very large data sets, a fourth data set was build from the combination of in-house compounds and compounds from external vendors. The number of molecules in this combined data set was 1 389 467.

As described in the methodology section, DISE requires three reference molecules for the sorting step. Although in theory arbitrary molecules or even fingerprints with random bits could be used, we decided to use fingerprints from drug molecules. Thus, molecules, which are more similar to the given drugs are selected first. The reference molecules used are shown in Figure 5. To make the reference molecules sufficiently dissimilar, the following procedure was applied: Starting from Sildenafil Citrate R1 the compounds from the MDDR database[22] were sorted by similarity and Isradipine (R2) was picked as second reference compound having a sufficient (tanimoto < 0.5) dissimilarity with R1. R3 was selected the same way making sure that it is dissimilar to R1 and R2.

RESULTS

Applying the Sphere Exclusion algorithm and DISE to the data sets described in the previous paragraphs yields the selections given in Table 1. Out of the 172 000 ASINEX compounds in set A, nearly 47 000 are selected by the Sphere Exclusion algorithm with a sphere radius of 0.15. The DISE algorithm selects more than 50 000 or 7.8% more compounds with the same sphere size. If singletons are disregarded (set B), the two algorithms select 30 000 and nearly 34 000 compounds, respectively. The difference here is more pronounced (11.5%). If only the more densely populated regions of this chemical space are considered (set C), the difference increases to 23%. This is because both methods will always select all singletons. It is only in densely populated regions of the candidate space that DISE has enough candidates to choose from to make a difference. Both

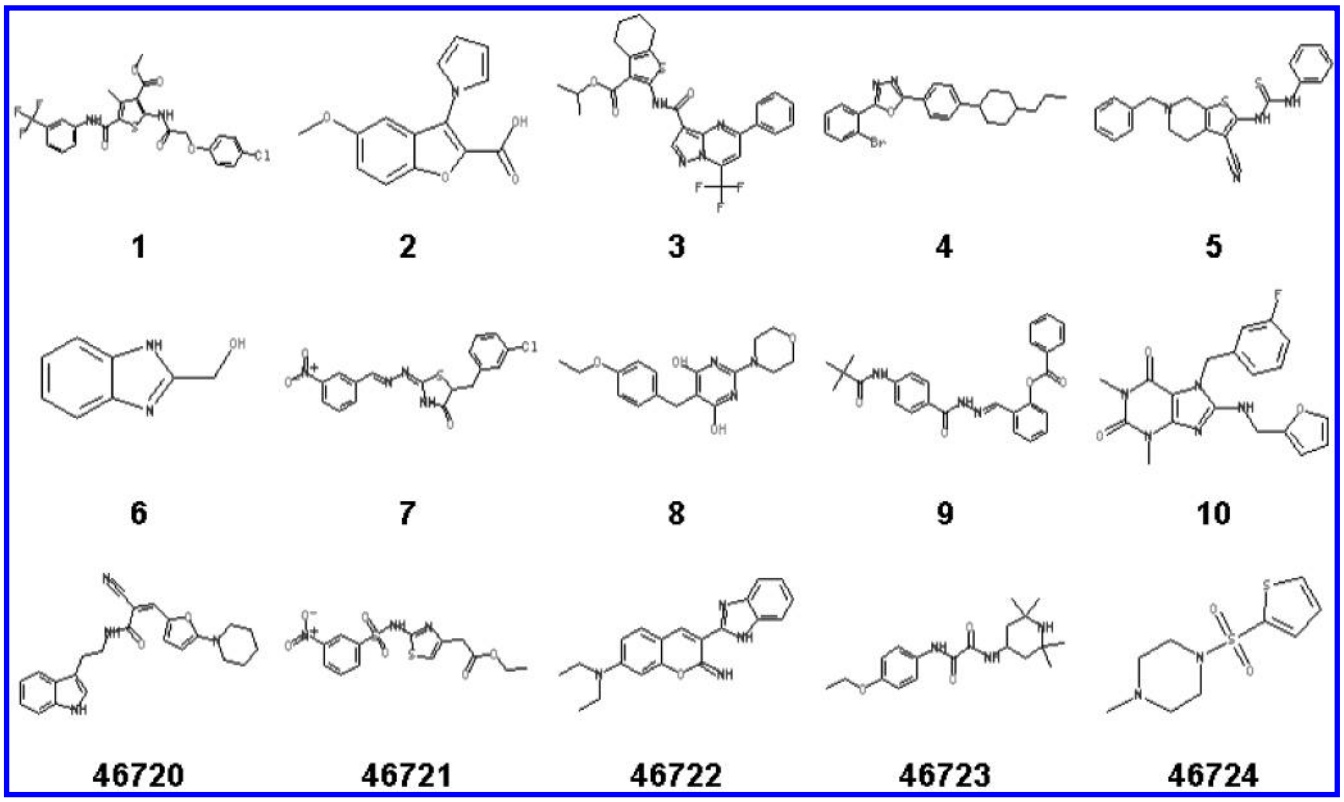

Figure 6. First and last compounds selected by the Sphere Exclusion algorithm applied to the complete ASINEX data set.

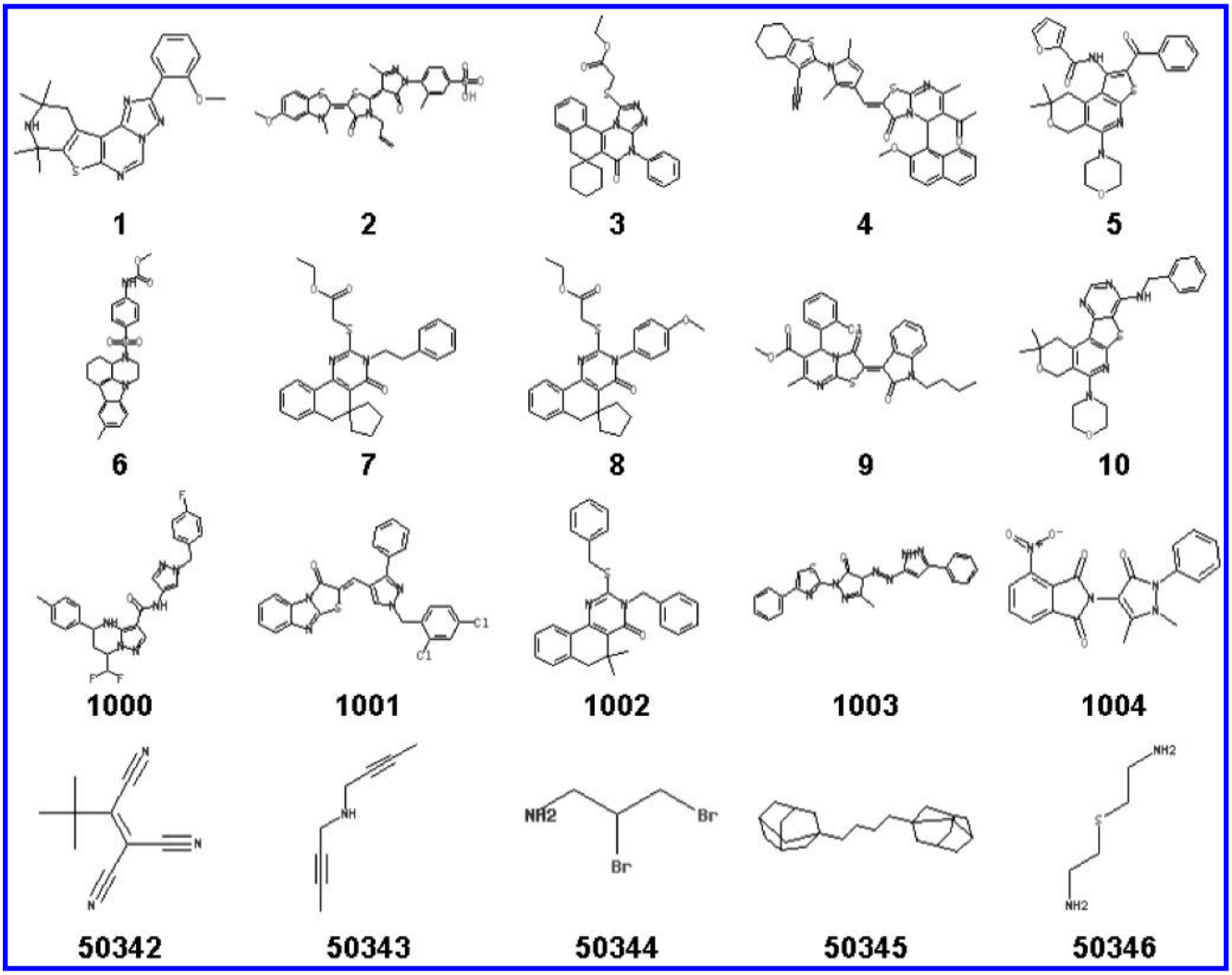

Figure 7. Compounds 1−10, 1000−1004 and last compounds as selected by the DISE algorithm applied to the complete ASINEX data set.

Table 1. Results of Sphere Exclusion and DISE Methods Applied to the Four Data Sets Described in the Text[a]

	set A	set B	set C	combined data set
compds	172 239	155 548	64 271	1 389 467
compds selected with SE	46 724	30 179	3759	308 965
compds selected with DISE	50 346	33 655	4635	332 886
% increase (compds)	7.8	11.5	23.3	7.7
CPU time [s] SE	3941	2278	86	196 234
CPU time [s] DISE	1968	969	24	76 386
% CPU time saved	50.1	57.5	72.0	61.1

[a] The radius of the spheres was set to 0.15.

methods select compounds from the same candidate pool to fill the space occupied by the candidate compounds with the same sphere size constraint. The higher number of selected compounds, therefore, suggests strongly that compounds selected by the DISE method are distributed more evenly in the predetermined chemical space. This suggestion is also supported by the data in Table 2 where the statistics of nearest neighbor distances between the compounds selected with both methods are given. The difference are not very pronounced for set B because there were too few candidates for DISE to make a big difference, however in set C the standard deviation of the nearest neighbor distance is 16% smaller using DISE than using the Sphere Exclusion method.

Table 2. Statistics on the Nearest Neighbor Distances of the Compounds Selected from Set B and Set C

		set B	set C
SE	average distance	0.188	0.175
	standard deviation	0.049	0.049
DISE	average distance	0.187	0.171
	standard deviation	0.046	0.041

This indicates clearly that the distribution of the compounds selected by DISE is more regular in comparison to those picked by the Sphere Exclusion method.

To check the dependency of the selection on the reference molecules, several different reference molecules were chosen and the ordering of the reference molecules was permutated. The results, in terms of number of compounds selected and CPU time, were essentially equal to the results presented in this paper.

Figure 6 shows the first and last compounds from the subset selected from data set A by the Sphere Exclusion algorithm. No systematic patterns can be found since the ordering of the compounds was random. The compounds shown in Figure 7 are those selected by the DISE algorithm. Due to the pre-sorting the first compound is the one with the highest similarity to the first reference compound R1 (Sildenafil Citrate). In this case, the similarity is 64%. The

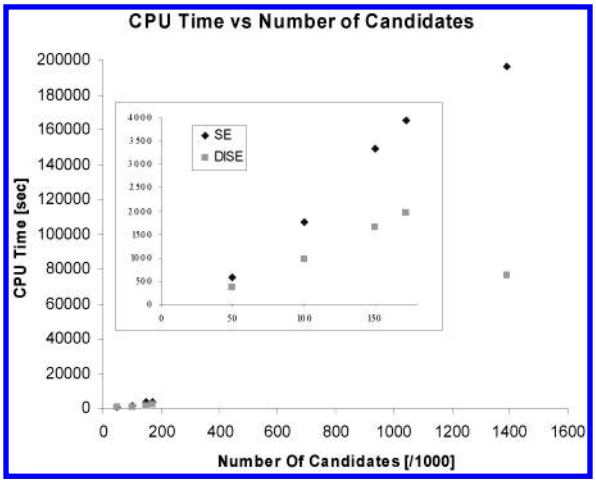

Figure 8. CPU time required for the compound selection using the Sphere Exclusion (♦) and the DISE (■) algorithms. The first 4 data points are for various random subsets of the ASINEX data set while the last point gives the times required for selection from the combined data set. The inset shows the first data points in more detail.

other compounds, which come right after the first one, have similarities to R1 in same range. All of them have features that exist in the first reference compound, e.g. the pyrimidine, pyrazole or sulfon fragments, and similarities to each other are visible. The pre-clustering effect of the sorting step is still visible for the compounds in the latter section of the sequence (1000−1004), although they are less similar to R1. The last compounds selected by the DISE algorithm are those, which are very dissimilar to the reference compounds. These are low molecular weight compounds or compounds such as 50 345, which have very few bits set in the daylight fingerprints. Some of them probably should have been excluded by a more sophisticated pre-filtering. However, it might be easier to let medicinal chemists check the last compounds after completing the selection, because those compounds will have the least similarity to the first reference compound and are therefore more likely to be of little interest. Compounds similar to these are also selected by the Sphere Exclusion method; they will however not be at the end of the selection list.

The significant performance benefit, which is achieved by the DISE algorithm as compared to the Sphere Exclusion method, is shown in Table 1 and Figure 8. It can be seen that the DISE algorithm yields 50−60% savings in CPU time. The savings are larger for the larger data sets, which means that the DISE algorithm is well suited for application to very large data sets. As discussed in the method section, the performance increase is achieved by first comparing each new candidate with the most recently selected compounds. Since the compounds were sorted first, many are discarded after fewer comparisons than using the SE method. To confirm this, we counted the number of comparisons needed to reject a compound in both algorithms when applied to the ASINEX data set. On average, the DISE algorithm needed 2222 comparisons to reject one compound while the Sphere Exclusion algorithm required 16 900 comparisons.

The exact dependence of the performance of DISE on the number of compounds is very much dependent on the composition of the candidate set. The order of the dependence of the algorithm on the number of compounds will vary between $O(N * n)$ and $O(n^2)$ where n is the number of compounds selected and N the total number of input compounds. The O(N x n) case will occur if all candidate compounds are singleton and none of them will be excluded. On the other hand if the sorting results in a "perfect" order the dependence would be of $O(n^2)$ because each rejected compound would be rejected after just one comparison and only the selected compounds are compared to all the other selected ones. The initial sorting step can be completed in O(N x ln(N)) steps and will always be much faster than the selection process.

SUMMARY

DISE, a modification of the Sphere Exclusion algorithm for selecting a diverse, evenly distributed subset of compounds for primary screening, was described and compared to the popular DIVPICT variant of the Sphere Exclusion algorithm. It was shown that DISE selects more compounds in a given chemical space implying a more even distribution of the selected compounds. In addition, DISE requires 50−60% less CPU time than the reference method. The performance of the DISE algorithm lends itself to be applied to data sets as large as multiple millions of compounds.

The results of the DISE algorithm share the same very intuitive geometric interpretation with the Sphere Exclusion method: each selected compound is separated by at least a sphere of given radius from all other selected compounds. If active compounds are found when screening the diverse subset, it is straightforward to go back to the full candidate set and select analogues by simple similarity search.

OUTLOOK

This paper presents compound selection for the primary screening as a case study to demonstrate the capabilities of the DISE algorithm. The "parameters", e.g. the sphere size and the choice of the reference molecules, were set accordingly. However, we envision that DISE is applicable to the compilation of subset for follow-up screens. In this case, one might use known ligands or previous hits as reference molecules and reduce the size of the exclusion sphere, to have compounds similar enough for thorough SAR studies. Depending on the scope, one might simply cut off the candidate compounds, which are too dissimilar to the reference molecules or one might even increase the sphere size during the run based on the distance of the candidates to the reference compound.

ACKNOWLEDGMENT

We would like to thank Robin Hewitt and Maya Binun for many helpful discussions and comments. We also would like to acknowledge the referees, which provided many useful hints to improve this paper.

REFERENCES AND NOTES

(1) Gillet, V. J.; Willett, P. Dissimilarity based compound selection for library design. In: Combinatorial library design and evaluation. Ghose, A. K.; Viswanadhan, V. N.; Eds.; Marcel Dekker Inc.: New York; 2001; pp 279−398.

(2) Eichler, U.; Ertl, P.; Gobbi, A.; Poppinger, D. Addressing the problem of molecular diversity. *Drugs Future* **1999**, *24*, 177−190.

(3) Martin Y.; A similarity investigation. Presentation at the MUG 02, 2002, Santa Fe, http://www.daylight.com/meetings/mug02/Martin/index.html.

(4) *Similarity and clustering in chemical information systems*; Willett, P., Ed.; Research Studies Press: Letchworth 1987.
(5) Downs, G. M.; Willett, P. Clustering of chemical structure databases for compound selection. *Methods Principles Med. Chem*. **1995**, *3* (Advanced computer assisted techniques in drug discovery), 111−130.
(6) Bayley, M. J.; Willett, P. Binning schemes for partition-based compound selection. *J. Mol. Graph Modeling* **1999**, *17*, 10−18.
(7) Pearlman R. S.; Smith K. M. Metric validation and the receptor relevant subspace concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(8) Pearlman, R. S. Novel Software tools for addressing chemical diversity. Network Science, 1996; http://www.netsci.org/Science/Combichem/feature08.html.
(9) Burden, F. R. Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225−227.
(10) Willett, P. Dissimilarity-based algorithms for selecting structural diverse sets of compounds. *J. Comput. Biol*. **1999**, *6*, 447−457.
(11) Lajiness, M.; Johnson, M. A.; Maggiora, G. M. Implementing drug screening using molecular similarity methods. In *QSAR: Quantitative Structure−Activity Relationships in Drug Design*; Fauchere, J. L., Ed.; Alan Liss, Inc.: New York, 1989; pp 173−176.
(12) 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.
(13) 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.
(14) Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graph Modeling* **1997**, *15*, 373−385.
(15) Nilakantan, R.; Bauman, N.; Haraki, K. S. Database diversity assessment: New ideas, concepts and tools. *J. Comput. Aided Mol. Design* **1997**, *11*, 447−452.
(16) Clark, R. D. OptiSim: An extended dissimilarity selection method for finding diverse representative subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181−1188.
(17) MacCuish, J.; Nicolaou, C.; MacCuish, N. E. Ties in proximity and clustering compounds. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 134−146.
(18) James, C. A. Weininger, D. Daylight Software Manual Version 4.42. Daylight Chemical Information Systems Inc., Irvine 1996; http://www.daylight.com.
(19) 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.
(20) 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*. **1998**, *39*, 3049−3059.
(21) ASINEX Ltd., 6 Schukinskaya St., Moscow 123182, Russia, http://www.asinex.com.
(22) MDL Drug Data Report, MDL Information Systems Inc. http://www.mdli.com/pdfs/MDDRds.pdf.

CI025554V