Published paper
Ashton, M., Barnard, J., Casset, F., Charlton, M., Downs, G., Gorse, D., Holliday, J., Lahana, R. and Willett, P. (2003) *Identification of diverse database subsets using property-based and fragment-based molecular descriptions,* Quantitative Structure-Activity Relationships, Volume 21 (6), 598 - 604.

Identification Of Diverse Database Subsets Using Property-Based And Fragment-Based Molecular Descriptions

Mark Ashton[a], John Barnard[b], Florence Casset[c], Michael Charlton[a], Geoffrey Downs[b], Dominique Gorse[c], John Holliday[d1], Roger Lahana[c] and Peter Willett[d]

[a] Evotec OAI, 151 Milton Park, Abingdon, Oxfordshire, OX14 4SD, UK
[b] Barnard Chemical Information, Ltd., 46 Uppergate Road, Sheffield S6 6BX, UK
[c] Synt:em, Parc Scientifique G. Besse, 30000 Nimes, France
[d] Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, UK

Keywords Diversity, Molecular diversity analysis, Structural diversity, Subset selection

Received on:

Abstract This paper reports a comparison of calculated molecular properties and of 2D fragment bit-strings when used for the selection of structurally diverse subsets of a file of 44295 compounds. MaxMin dissimilarity-based selection and *k*-means cluster-based selection are used to select subsets containing between 1% and 20% of the file. Investigation of the numbers of bioactive molecules in the selected subsets suggest: that the MaxMin subsets are noticeably superior to the *k*-means subsets; that the property-based descriptors are marginally superior to the fragment-based descriptors; and that both approaches are noticeably superior to random selection.

INTRODUCTION

Developments in combinatorial chemistry and high-throughput screening over the last few years mean that it is now possible to synthesise and test far greater numbers of compounds in lead discovery programmes than was previously possible. However, molecules that are structurally similar are likely to exhibit comparable activity profiles, and considerations of cost-effectiveness hence dictate that the compounds chosen for inclusion in such programmes should be structurally diverse, so as to maximise the amounts of structure-activity information that can be obtained without redundant experimentation. The need for structural diversity has resulted in the development of computer-based methods for selecting libraries of molecules that ensure coverage of the largest possible expanse of chemical space in the search for new leads, an area of study that is normally referred to as *molecular diversity analysis*

[1] To whom all correspondence should be addressed.

[1-4]. There is already an extensive literature associated with many aspects of compound selection covering topics such as representations to describe the molecules for diversity analysis [5], ways of quantifying the degree of structural (dis)similarity between pairs of molecules [6], algorithms for identifying the presence of distinct clusters of molecules [7] and for identifying structurally disparate database subsets [8], and indices to quantify the diversity of such subsets [9]. In this paper, we focus on the first of these aspects, *viz* the structural representations that are used to characterise molecules. Specifically, we report a comparison of two of the most common types of description, these being calculated physical properties and 2D fragment bit-strings.

Several comparative studies of methods for diversity analysis have been reported previously (see, e.g., [10-17]) but this study has two characteristics of particular interest. First, it involves compounds synthesised and tested as part of the ongoing commercial operations of Evotec OAI (hereafter EOAI) [18], rather than a standard public database such as the *World Drug Index* or *MACCS Drug Data Report* files. Second, as the aim was to compare two representative, but rather different, operational software systems, a detailed experimental design was required to ensure the elimination of other features of a diversity analysis and to ensure that the comparison focused just upon the particular types of representation studied here. These representations were the physical property descriptions used in the Diverser software produced by Synt:em [19-21] and the 2D fragment bit-strings used in the Barnard Chemical Information (hereafter BCI) software [22-24]: in what follows, we shall refer to these as *property-based* (PB) and *fragment-based* (FB) descriptors.

MATERIALS AND METHODS

Dataset

The dataset provided by EOAI contains 44295 compounds, each of which had an associated biological activity value obtained from a single assay on one specific biological target. The 4505 molecules with a percentage inhibition of >40% are regarded as hits by EOAI: in what follows, these will be referred to as 'Actives'. Two subsets of these compounds were also defined for the purposes of this comparison: the 2750 'Moderately-High Actives' have inhibitions of >60% and the 1656 'High Actives' have inhibitions of >80%.

Structure Representations

The Diverser software calculates a large number of topological, property and shape descriptors for each of the molecules in a dataset, specifically those generated by the Molconn-Z software system [25] and then employs a variable-elimination procedure to identify a small number of discriminating descriptors. Here, a total of 327 descriptors was initially defined for each of the molecules in the dataset. Of these, 45 were removed as exhibiting null variance and 129 as exhibiting inappropriate frequency distributions. The remaining 153 descriptors then underwent a correlation analysis, which eliminated a further 66 variables having correlations ≥ 0.95 with descriptors chosen for retention. The resulting set of 87 statistically significant descriptors was then subjected to principal component analysis, yielding a set of 16 principal components that explained 99.6% of the variance in the original dataset. These components formed the structural description for each of the molecules in the

dataset; specifically, each of the 16 components was encoded in several bits (the precise number for each being determined by the percentage of the variance explained) and molecular bit-strings finally obtained by concatenating the pattern of bits for each component. The encoding of the features is discussed by Gorse *et al.* [21].

This PB representation of molecular structure was compared with an FB representation, specifically the 2D substructural features that are encoded in the bit-strings generated by the BCI software; similar sets of features are used in other commercial packages for chemical information management produced by organisations such as Daylight Chemical Information Systems Inc., MDL Information Systems Inc. and Tripos Inc. The BCI bit-strings encode the presence or absence in molecules of 2D substructural fragments from an externally-defined fragment dictionary; each bit position is directly associated with a particular fragment, defined in the dictionary (in principle, a group of related fragments can be associated with a single bit position, though this was not done here.) BCI provides general-purpose fragment dictionaries, but normally expects better results to be obtained from a custom dictionary that reflects the substructural occurrences and co-occurrences in the specific dataset that is to be processed, and this was found to be the case here. All of the results reported below are based on the use of a custom dictionary, which was generated as follows. First, the EOAI dataset was processed to identify all fragments occurring in it in the following five families: Augmented Atom (an atom and its immediate neighbours), Atom Sequence (linear atom-bond paths of between 3 and 6 atoms), Atom Pair (pairs of atoms with the topological distance between them, Ring Composition (atom sequences around individual rings) and Ring Fusion (describing the fusion patterns in multicyclic systems). After initial identification, each fragment was generalised by replacing the specific atom and bond types by generalised values, using intermediate types such as “halogen” and “ring bond”, and fully-generalised “any atom” and “any bond” types. This yielded a total of over 21,000 distinct fragments, at specific and generalised levels, which were then reduced by eliminating those that occurred in less than 10% or more than 25% of the molecules, or whose frequency of occurrence was too close to that of another, less specific fragment of the same family. These eliminations were intended to remove fragments that provided little differentiation between the molecules, and to replace groups of less-common fragments by their common generalisations, while avoiding the inclusion of redundant specific and generalised descriptions of the same features. The fragment selection process requires manual intervention to choose cut-off frequencies etc., and the choices made were aimed primarily at producing bit-strings of a size comparable to those commonly used in diversity analysis work. The process (at least in its present implementation) is somewhat laborious, with many user-definable options: the final dictionary used here contained a total of 1073 fragments selected from the initial set of over 21,000 fragments. The encoding of the features is described on the BCI Web site [22].

Subset Selection

Two approaches were adopted for subset selection that are available in both the Diverser and BCI software: MaxMin dissimilarity-based selection and *k*-means cluster-based selection.

MaxMin is perhaps the most widely used method for dissimilarity-based compound selection. Assume that a *Subset* of *k* molecules is to be selected from a *Dataset* containing *N* molecules. Then the MaxMin method is as follows:

1. Initialise *Subset* with some appropriately chosen seed compound and set x:=1.
2. For each of the N-x remaining compounds in *Dataset* calculate its dissimilarity with each of the x compounds in *Subset* and retain the smallest of these x dissimilarities for each compound in *Dataset*.
3. Select the molecule from *Dataset* with the largest value for the smallest dissimilarity calculated in Step 2 and transfer it to *Subset*.
4. Set x:=x+1 and return to Step 2 if $x < k$.

Cluster-based selection provides an alternative way of identifying a structurally diverse database subset. The *k*-means method was used here, as follows:

1. A set of *k* initial cluster representatives is selected (where *k* is the number of clusters required) and the centroids calculated.
2. Each molecule in *Dataset* is assigned to the cluster with the closest representative.
3. New representatives are calculated for each cluster, reflecting the assignments made in Step 2, and Step 2 is repeated if the clustering has not converged.
4. *Subset* is then the molecules comprising the final set of cluster representatives.

Care was taken to ensure the removal of possible sources of variation in the two software implementations, so that any differences observed could be assumed to result from the representations employed. Factors considered included the following. First, MaxMin requires a starting-point compound: that chosen was the compound closest to the centre of the dataset, where centre was defined as the arithmetic mean of the representations of the whole dataset. Second, *k*-means requires the specification of cluster representatives that are updated at the end of each complete pass through the dataset, if order-independent subsets are to be obtained: the representative for a cluster used here was the individual compound nearest to the arithmetic mean of the representations of the compounds in that cluster, with this nearest compound at the end of the clustering being the compound that was chosen for inclusion in the selected subset. *k*-means also requires the specification of an initial set of cluster representatives: those used here were the compounds resulting from the MaxMin selections. Finally, both systems used Euclidean distance as the similarity coefficient, with some of the FB experiments using the Soergel distance (the complement of the Tanimoto coefficient) in the nearest neighbour experiments described below. These two distance coefficients are discussed by Willett *et al*. [6] and are detailed in Figure 1.

Subset Evaluation

Subsets were generated, using the two approaches above, that contained 1, 2, 3, 5, 10 and 20% of the database, these corresponding to k = 443, 886, 1329, 1772, 2215, 4430 and 8860 molecules, respectively. The subsets were then analysed to determine the effectiveness of the representations that had given rise to them.

A common procedure in drug discovery is to take an initial set of actives (identified by whatever means) and then to use these to retrieve further compounds that have high probabilities of activity. Indeed, the Diverser software uses the actives to build sophisticated filters that can then be used for the screening of previously untested molecules. A simpler approach, and the one used here as it is feasible in both of the

software systems being compared, involves a *feedback* or *expansion* experiment, which involves taking actives that have been identified in a subset and then using these to retrieve further compounds that are expected to exhibit the same activity.

Assume that a classification has been made of a dataset (this is done here using the *k*-means clustering method) so that a subset can be obtained by selecting one or more molecules from each of the clusters in the classification. Having identified the actives in the subset, and hence the *active clusters* from which they are derived, one can then obtain feedback molecules by considering the other molecules in each of the active clusters (as the best classification is presumably one in which the actives are maximally clustered together, with the active clusters containing as few inactives as possible). Thus the expansion set here is the compounds that are in the active clusters that have been identified. With a MaxMin-derived subset, expansion is achieved by taking molecules similar to the actives in the chosen subset: it was decided to take the 10 nearest neighbours (NNs), i.e., the most similar compounds, so that the expansion set here was the NNs of each of the initial set of actives. The NNs were identified using the Euclidean distance (PB) and using the Euclidean distance and the Soergel distance (FB). These coefficients are The latter was found to give slightly, but consistently, better results than the Euclidean distance; we hence describe only the Soergel NN experiments when discussing the FB results below.

RESULTS

The PB MaxMin results are detailed in Table 1 and the FB MaxMin results in Table 2; the corresponding sets of *k*-means results are detailed in Tables 3 and 4, respectively. The pairs of columns in Tables 1 and 2 represent the number of actives in the MaxMin subset and the mean percentage of actives in the sets of 10 NNs for each of the actives in the MaxMin subset. The pairs of columns in Tables 3 and 4 represent the number of active clusters and the mean percentage of actives *per* active cluster.

DISCUSSION

Comparison Of MaxMin Selections

A comparison of Tables 1 and 2 leads to two general conclusions. First, looking at the "No. Actives" columns, the PB subsets nearly always contain a larger number of active molecules (using any of the three definitions of activity) than do the FB subsets. Second, looking at the "Ave. % NN Actives" columns, the FB feedback subsets normally, but not always, contain a larger number of active molecules (using any of the three definitions of activity) than do the PB feedback subsets.

These differences are not large but they do apply across the results presented in the tables. For example, considering all of the active molecules (low, medium and high), the percentage difference in the number of actives, i.e.,

$$\frac{\left|Actives(PB) - Actives(SB)\right|}{Min\{Actives(PB), Actives(SB)\}} \times 100,$$

ranges from 1.6% (for the 443-molecule subset) to 11.2% (for the 1772-molecule subset), with a median percentage difference of 5.9%. Again considering all of the actives, the percentage difference in the number of NN actives, i.e.,

$$\frac{\left|ActivesNN(PB) - NNActives(SB)\right|}{Min\{ActivesNN(PB), ActivesNN(SB)\}} \times 100$$

ranges from 0 (for the 443-molecule subset) to 23.3% (for the 4430- and 8860-molecule subsets), with a median percentage difference of 15.1%.

One may thus conclude that the PB initial subsets are a richer source of actives than the FB initial subsets, but that the NNs of the latter sets of molecules are likely to produce more actives in the feedback stage.

Comparison Of *k*-Means Selections

A comparison of Tables 3 and 4 leads to two general conclusions. First, looking at the "No. Active Clusters" columns, there are no consistent differences and it would be difficult to argue that one approach was noticeably superior to the other. Thus, PB yields more active clusters with small and large subsets or when just the High Actives are considered, but FB is the better with the intermediately-sized subsets. Second, looking at the "Ave. % Actives in Active Clusters" columns, the PB feedback clusters normally, but not always, have a greater percentage of active compounds in the active clusters than do the FB active clusters.

Considering the percentage differences (calculated as described above but using the cluster representatives rather than the MaxMin molecules) for all of the active molecules then the differences range from 1.4% (FB doing better with the 1329-molecule subset) to 27.9% (PB doing better with the 443-molecule subset), with a median difference of 11.1% (BCI doing better with the 1772-molecule subsets).

The differences between the two approaches can, rather crudely, be summarised as shown in Table 5. Here, we have simply considered how many times the PB (or FB) result was better than the FB (or PB) result, when summed over all of the entries in Tables 1 and 2 and in Tables 3 and 4. The crudeness arises from the fact that the entries in Tables 1-4 are not independent of each other: High Actives $\subseteq$ Mod.-High Actives $\subseteq$ Actives and subset-433 $\subseteq$ subset-886 *etc*. (so that if the first subset is poor then subsequent ones are also likely to perform badly). None the less, the table does provide a rapid summary of the trends noted above from the MaxMin and *k*-means experiments.

There has been considerable interest in the evaluation of different types of representation for use in diversity analyses, with several of the studies suggesting that 2D fragment bit-strings provide a generally high level of performance [5, 7, 10-12]. These studies have generally used public datasets. Our results, obtained with a large operational file, demonstrate that property-based molecular descriptors can provide equally effective representations of molecular structure if appropriate variable-selection and encoding methods are used.

Comparison Of MaxMin and *k*-Means Selections

One conclusion that can be drawn from comparing Tables 1 and 2 with Tables 3 and 4 is that the MaxMin subsets normally, but not always, contain a larger number of

active molecules than do the *k*-means subsets. Moreover, in the case of the FB feedback subsets, the sets of top-10 NNs have a rather larger percentage of actives than do the molecules in the active clusters from the *k*-means analysis; there does not seem to be such an obvious difference in the case of the PB feedback results. It is not clear why there is such a discrepancy between the MaxMin and *k*-means feedback results, as the initial subsets from the former procedure provided the seeds for the clustering experiments. This might possibly be related to the "natural" number of clusters in the dataset: the Average % Actives in Active Clusters columns in Tables 3 and 4 certainly suggest that a plateau of performance is reached around 1772 molecules, whereas no such trend is seen with the MaxMin results.

As the MaxMin results are generally superior to the *k*-means results, we have carried out a further comparison of the two sets of feedback compounds. Specifically, Table 6 lists the average percentage of actives in the NNs of actives molecules from the initial MaxMin subsets and the average percentage of actives in the NNs of inactive molecules from the initial MaxMin subsets. It will be seen that there is a very well-marked concentration of actives around the actives in the initial subsets, as against the inactives in the initial subsets: we hence deduce that only a small number of active molecules are lost by feedback experiments that consider the NNs of just the initial actives.

Comparison With Random Selection

The last comparison to be reported is with the numbers of actives (all classes of actives are considered here) that would be expected in subsets obtained by random selection, these expected numbers being calculated from the total number of actives in the dataset and from the subset sizes. This comparison is summarised in Table 7, where it will be seen that with two exceptions (one PB and one FB) the systematic subsets contain more active molecules than would be expected by simple random selection.

Conclusions

In this paper we have compared the use of property-based and structure-based molecular representations for molecular diversity analysis. Our results suggest that the MaxMin method for dissimilarity-based selection is superior to the *k*-means method for cluster-based selection. Of the two representations, the property-based descriptors generally permit the generation of subsets that contain a larger number of active molecules than do the fragment-based descriptors; that said, some of the differences in performance are very small and there are many cases where the fragment-based subsets are to be preferred. Finally, and perhaps most importantly, comparison with random selection demonstrates that both approaches provide an appropriate way of selecting molecules from computer databases that could be used in an operational context for synthesis and/or biological testing.

Acknowledgements We thank Barnard Chemical Information, Evotec OAI, the Royal Society, Synt:em and the Wolfson Foundation for funding and software support. The Krebs Institute for Biomolecular Research is a designated Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

REFERENCES

1. Special issue on "Computational Methods for the Analysis of Molecular Diversity." *Perspect. Drug Discov. Design 7/8*, 1-180 (1997).
2. Dean, P.M. and Lewis, R.A., *Molecular Diversity in Drug Design*, Kluwer, Amsterdam 1999.
3. Special issue on "Combinatorial Library Design" *J. Mol. Graph. Model. 18*, 317-540 (2000).
4. Ghose A.K. and Viswanadhan, V.N., *Combinatorial Library Design and Evaluation: Principles, Software Tools and Applications in Drug Discovery*, Marcel Dekker, New York 2001.
5. Brown, R.D., Descriptors for Diversity Analysis, *Perspect. Drug Discov. Design 7/8*, 31-49 (1997).
6. Willett P., Barnard, J.M. and Downs, G.M., Chemical Similarity Searching, *J. Chem. Inf. Comput. Sci. 38*, 983-996 (1998).
7. Wild, D. and Blankley, C. J., Comparison of 2D Fingerprint Types and Hierarchy Level Selection Methods for Structural Grouping Using Ward's Clustering, *J. Chem. Inf. Comput. Sci. 40*, 155-162 (2000).
8. Higgs, R.E., Bemis, K.G., Watson, I.A. and Wikel, J.H., Experimental Designs for Selecting Molecules from Large Experimental Databases, *J. Chem. Inf. Comput. Sci. 37*, 861-870 (1997).
9. Waldman, M., Li, H. and Hassan, M., Novel Algorithms for the Optimization of Molecular Diversity of Combinatorial Libraries, *J. Mol. Graph. Model. 18*, 412-426 (2000).
10. Brown, R.D. and 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. 36*, 572-584 (1996).
11. Patterson D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D. and Weinberger, L.E., Neighbourhood Behaviour: a Useful Concept for Validation of "Molecular Diversity" Descriptors, *J. Med. Chem. 39*, 3049-3059 (1996).
12. Matter, H., Selecting Optimally Diverse Compounds from Structural Databases: a Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors, *J. Med. Chem. 40*, 1219-1229 (1997).
13. Snarey, M., Terret, N.K., Willett, P. and Wilton, D.J., Comparison of Algorithms for Dissimilarity-Based Compound Selection, *J. Mol. Graph. Model. 15*, 372-385 (1997).
14. Matter, H. and Potter, T., Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets, *J. Chem. Inf. Comput. Sci. 39*, 1211-1225 (1999)
15. Bayada D.M., Hamersma H. and van Geerestein, V.J., Molecular Diversity and Representativity in Chemical Databases, *J. Chem. Inf. Comput. Sci. 39*, 1-10 (1999)
16. Gute, B.D., Grunwald, G.D., Mills, D. and Basak, S.C., Molecular Similarity Based Estimation of Properties: a Comparison of Structure Spaces and Property Spaces, *SAR QSAR Environ. Res. 11*, 363-382 (2001).
17. Andersson, P.M., Sjostrom, M., Wold, S. and Lundstedt, T., Strategies for Subset Selection of Parts of an In-House Chemical Library, *J. Chemomet. 15*, 353-369 (2001).
18. Evotec OAI is at URL http://www.evotecoai.com/
19. Synt:em is at URL http://www.syntem.com
20. Grassy, G., Calas, B., Yasri, A., Lahana, R., Woo J., Iyer, S., Kaczorek, M., Floc'h and Buelow, R., Computer-Assisted Rational Design of Immunosuppresive Compounds, *Nature Biotech. 16*, 748-752 (1998).
21. Gorse, D., Rees, A., Kaczorek, M. and Lahana, R., Molecular Diversity and its Analysis, *Drug Discov. Today 4*, 257-264 (1999).
22. Barnard Chemical Information is at URL http://www.bci1.demon.co.uk/
23. Barnard, J.M. and Downs, G.M., Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures, *J. Chem. Inf. Comput. Sci. 32*, 644-649 (1992).
24. Barnard, J.M. and Downs, G.M., Chemical Fragment Generation and Clustering Software, *J. Chem. Inf. Comput. Sci. 37*, 141-142 (1997).
25. The Molconn-Z software is available from eduSoft LC at URL http://www.edusoft-lc.com/

Subset	Actives		Mod-High Actives		High Actives	
	No. Actives	Ave. % NN Actives	No. Actives	Ave. % NN Actives	No. Actives	Ave. % NN Actives
443	62	37	42	35	25	36
886	125	33	90	31	53	29
1329	181	32	123	30	72	28
1772	238	33	164	32	100	31
2215	290	33	194	31	119	29
4430	547	30	345	28	205	26
8860	1080	30	657	29	391	26

Table 1: Property-based MaxMin selection

Subset	Actives		Mod-High Actives		High Actives	
	No. Actives	Ave. % NN Actives	No. Actives	Ave. % NN Actives	No. Actives	Ave. % NN Actives
443	61	37	39	36	25	30
886	118	37	76	36	39	28
1329	176	37	109	37	60	28
1772	214	37	135	37	77	28
2215	268	38	169	38	103	30
4430	525	37	328	36	200	30
8860	1009	37	633	36	375	31

Table 2: Fragment-based MaxMin selection

Subset	Actives		Mod-High Actives		High Actives	
	No. Active Clusters	Ave. % Actives in Active Clusters	No. Active Clusters	Ave. % Actives in Active Clusters	No. Active Clusters	Ave. % Actives in Active Clusters
443	55	29	32	28	24	24
886	107	35	62	37	43	27
1329	147	36	94	37	67	28
1772	188	36	119	37	82	30
2215	225	35	144	35	95	27
4430	463	34	286	35	187	29
8860	993	28	587	29	366	25

Table 3: Property-based k-means selection

Subset	Actives		Mod-High Actives		Highly Active	
	No. Active Clusters	Ave. %. Actives in Active Clusters	No. Active Clusters	Ave. % Actives in Active Clusters	No. Active Clusters	Ave. % Actives in Active Clusters
443	43	31	23	31	18	22
886	93	33	55	32	39	28
1329	149	35	89	36	62	30
1772	209	35	133	36	93	28
2215	251	33	153	33	95	29
4430	486	28	297	27	181	23
8860	919	23	560	22	343	19

Table 4: Fragment-based k-means selection

Approach	MaxMin		*k*-Means	
	No. Actives	Ave. % NN Actives	No. Active Clusters	Ave. % Actives in Active Clusters
Property-Based	20	3	12	16
Fragment-Based	0	16	8	5

Table 5: Numbers of times that each approach was superior to the other

Subset	Property-Based		Fragment-Based	
	Ave. %. Actives in NN Actives	Ave. %. Actives in NN Inactives	Ave. %. Actives in NN Actives	Ave. % Actives in NN Inactives
443	37	8	37	7
886	33	8	37	7
1329	32	8	37	7
1772	33	8	37	8
2215	33	8	38	8
4430	30	8	37	8
8860	30	8	37	8

Table 6: Percentage of actives in feedback sets based on MaxMin initial actives and inactives

Subset	Random	Property-Based		Fragment-Based	
		MaxMin	*k*-Means	MaxMin	*k*-Means
443	45	62 (+38%)	55 (+22%)	61 (+36%)	43 (-4%)
886	90	125 (+39%)	107 (+19%)	118 (+31%)	93 (+3%)
1329	135	181 (+34%)	147 (+9%)	176 (+30%)	149 (+10%)
1772	180	238 (+32%)	188 (+4%)	214 (+19%)	209 (+16%)
2215	225	290 (+29%)	225 (0%)	268 (+19%)	251 (+12%)
4430	451	547 (+19%)	463 (+3%)	525 (+16%)	486 (+8%)
8860	901	1080 (+20%)	993 (+10%)	1009 (+12%)	919 (+2%)

Table 7: Numbers of actives in random, property-based and fragment-based selections

Figure 1: The Euclidean distance and the Soergel distance. In the case of continuous attributes, let the molecules A and B be represented by vectors such that x_{jA} is the value of the j-th attribute ($1 \le j \le n$, the total number of distinct attributes) in molecule A (and similarly for molecule B). In the case of binary attributes, a, b and c denote the numbers of bits set to "on" in A, in B, and in both A and B, respectively.

Formula	Euclidean Distance	Soergel Distance
Continuous variables	$D_{A,B} = \sqrt{\sum_{j=1}^{j=n} \left(x_{jA} - x_{jB} \right)^2}$	$D_{A,B} = \frac{\sum_{j=1}^{j=n} \left\| x_{jA} - x_{jB} \right\|}{\sum_{j=1}^{j=n} \max\left(x_{jA}, x_{jB} \right)}$
Binary variables	$D_{A,B} = \sqrt{a + b - 2c}$	$D_{A,B} = 1 - \frac{c}{a + b - c} = \frac{a + b - 2c}{a + b - c}$