Published paper

Willett, P. (2005) Evaluation of Molecular Similarity and Molecular Diversity Methods Using Biological Activity Data in *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*, Bajorath (ed.) Humana Press: New York pp.51-63

Contribution to *Methods in Molecular Biology*

Full title: The Evaluation Of Molecular Similarity And Molecular Diversity Methods Using Biological Activity Data

Running title: Evaluation Of Similarity And Diversity Methods

Author: Peter Willett

Degrees: MA, MSc, PhD, DSc

Affiliation and contact details: Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, UK. Email: p.willett@sheffield.ac.uk. Phone: +44-114-2222633. Fax: +44-114-2780300

Abstract. This paper reviews the techniques available for quantifying the effectiveness of methods for molecule similarity and molecular diversity, focusing in particular on similarity searching and on compound selection procedures. The evaluation criteria considered are based on biological activity data, both qualitative and quantitative, with rather different criteria needing to be used depending on the type of data available.

Keywords. Chemical database; Compound selection; Library design; Molecular diversity; Molecular similarity; Neighbourhood behaviour; Similar property principle; Similarity searching

1. Introduction

The concepts of molecular similarity ***(1-3)*** and molecular diversity ***(4, 5)*** play important roles in modern approaches to computer-aided molecular design. Molecular similarity provides the simplest, and most widely used, method for virtual screening and underlies the use of clustering methods on chemical databases. Molecular diversity analysis provides a range of tools for exploring the extent to which a set of molecules spans structural space, and underlies many approaches to compound selection and to the design of combinatorial libraries. Many different similarity and diversity methods have been described in the literature, and new methods continue to appear. This raises the question of how one can compare different methods, so as to identify the most appropriate method(s) for some particular application: this paper provides an overview of the ways in which this can be carried out, illustrating such comparisons by, principally, our experience of similarity and diversity studies that have been carried out in the Chemoinformatics Research Group at the University of Sheffield.

There are two bases for the comparison of similarity and diversity methods. It is possible to compare the *efficiency* of methods, i.e., the resources, typically computer time and computer memory, necessary for the completion of processing. Considerations of efficiency, in particular theoretical analyses of computational complexity, are important in that they can serve to identify methods that are unlikely to be applicable given the rapidly increasing sizes of current and planned chemical datasets. Here, however, we restrict ourselves to comparing the *effectiveness* of similarity and diversity methods, i.e., the extent to which a method is able to satisfy the user’s requirements in terms of identifying similar or diverse sets of compounds. More specifically, we focus on evaluation criteria based on the availability of bioactivity data for the molecules that are being processed, where the data can either be *qualitative*, i.e., a categorical (usually binary) variable, or *quantitative*, i.e., a real-valued variable. The discussion here considers only the criteria that can be used for comparative studies: the reader is referred elsewhere for the results of such studies.

2. Methods

2.1. Molecular Similarity Methods

2.1.1. Introduction

The basic concept of molecular similarity has many applications ***(1,2)*** but we focus here on its use for similarity-based virtual screening, which is often referred to as *similarity searching* ***(3)***. Here, a user specifies a *target structure* that is characterised by one or more structural descriptors, and this set is compared with the corresponding sets of descriptors for each of the molecules in the database. These comparisons enable the calculation of a measure of similarity, i.e., the degree of structural relatedness, between the target structure and each of the database structures, and the latter are then sorted into order of decreasing similarity with the target. The output from the search is a ranked list in which the structures that are calculated to be most similar to the target structure, the *nearest neighbours*, are located at the top of the list. These neighbours form the initial output of the search and will be those that have the greatest probability of being of interest to the user, given an appropriate measure of inter-molecular structural similarity.

Many different types of similarity measure have been discussed in the literature but they generally involve three principal components: the *representation* that is used to characterise the molecules that are being compared; the *weighting scheme* that is used to assign differing degrees of importance to the various components of these representations; and the *similarity coefficient* that is used to provide a quantitative measure of the degree of structural relatedness between a pair of structural representations. These three components are closely related and it is hence most important that a comparative study should seek to ensure that only one of these components is varied at any one time. For example, only a limited amount of information might be gained from a comparison of the effectiveness of similarity searching using binary fingerprints (e.g., those produced by the UNITY or Daylight software) and the Tanimoto coefficient, with the effectiveness of similarity searching using a set of computed physicochemical parameters (e.g., those produced by the MOLCONN-Z or DiverseSolutions software), some particular standardisation method and the Euclidean distance. Given an appropriate evaluation criterion (as discussed below), one might be able to decide that one of these approaches gave better results than the other, but one would not be able to identify the relative contributions of the various components of the overall similarity measures that were being studied.

The basis for all of the evaluation techniques to be discussed here is what is commonly referred to as the *similar-property principle*, which was first stated explicitly by Johnson and Maggiora in their seminal 1990 book ***(1)***. The principle states that structurally-similar molecules are expected to exhibit similar properties. It is clear that there are many exceptions to the principle as stated ***(6,7)***, since even a small change in the structure of a molecule can bring about a radical change in some property; for example, replacement of a small alkyl group by a larger one, e.g., methyl replaced by t-butyl, can mean that a molecule is now too large to fit a binding site. The principle does, however, provide a general rule of thumb that is very widely applicable; indeed, if this were not the case, then it would prove difficult indeed to develop meaningful structure-activity relationships of any sort. If the principle does hold for a particular dataset, then the top-ranked molecules (which are often referred to as the *nearest neighbours*) in a similarity search are expected to have properties that are related to those of the target structure. We can hence evaluate the effectiveness of a structurally-based similarity procedure by the extent to which the similarities resulting from its use mirror similarities in some external property, which in the context of this paper we take to be biological activity (but could be any type of chemical, biological or physical property). The next two sections of the paper detail the ways in which the principle is applied to the analysis of qualitative and quantitative datasets.

2.1.2. Use of qualitative data

In what follows, we shall adopt ideas and terminology from that part of computer science that is normally referred to as *information retrieval* ***(8-10)***. The measurement of search effectiveness has played a large part in the development of information retrieval (or IR) systems, whose principal aim is to identify as many documents as possible that are relevant to a user's query whilst simultaneously minimising the number of non-relevant documents that are retrieved. It is possible to apply many of these measures to the evaluation of chemical retrieval systems, where one wishes to identify as many molecules as possible that have the same activity as the target structure whilst simultaneously minimising the number of inactive molecules that are retrieved.

The relationship between IR and chemical similarity searching is discussed in detail by Edgar *et al.* ***(11)*** who summarise the various effectiveness measures in terms of the 2×2 contingency table shown in Table 1. In this table, it is assumed that a search has been carried out resulting in the retrieval of the *n* nearest neighbours at the top of the ranked output. Assume that these *n* nearest neighbours include *a* of the *A* active molecules in the complete database, which contains a total of *N* molecules. Then the ***recall***, *R*, is defined to be the fraction of the active molecules that are retrieved, *i.e.*,

$$R = \frac{a}{A},$$

and the ***precision***, *P*, is defined to be the fraction of the retrieved molecules that are active, *i.e.*,

$$P = \frac{a}{n}.$$

A retrieval mechanism should seek to maximise both the recall and the precision of a search so that, in the ideal case, a user would be presented with all of the actives in the database without any additional inactives: needless to say, this ideal is very rarely achieved in practice.

It is inconvenient to have to specify two measures, i.e., recall and precision, to quantify the effectiveness of a search. The Merck group have made extensive use of the ***enrichment factor***, i.e., the number of actives retrieved relative to the number that would have been retrieved if compounds had been picked from the database at random ***(12).*** Thus, using the notation of Table 1, the enrichment factor at some point, *n*, in the ranking resulting from a similarity search is given by

$$\frac{a/n}{A/N}.$$

Note that since *A*/*N* is a constant, the enrichment is monotonic with precision. Rather than specifying the enrichment at some specific point in the ranking, e.g., the top-1000 positions, it can alternatively be specified at that point where some specific fraction, e.g., 50%, of the actives have been retrieved. Examples of the use of enrichment factors are provided by Sheridan *et al*. ***(12)*** and Gillet *et al*. ***(13)***.

Alternatively, Güner and Henry ***(14)*** have introduced the G-H score, which is a weighted average of recall and precision. The score was originally developed for evaluating the effectiveness of 3D database searches but can be applied to the evaluation of any sort of search for which qualitative bioactivity data are available. Using the previous notation, the G-H score is defined to be

$$\frac{aP + bR}{2},$$

where α and β are weights describing the relative importance of recall and precision. The lowerbound for the G-H score is zero; if both weights are set to unity, then the score is simply the arithmetic mean of recall and precision, i.e.,

$$\frac{P+R}{2}.$$

Examples of the use of the G-H score are provided by Güner and Henry ***(15)*** and by Raymond and Willett ***(16)***, while Edgar *et al*. discuss other combined measures that can be used for chemical similarity searching ***(11).***

At least three alternative approaches have been widely used. First, the Sheffield group has generally quoted the mean numbers of active compounds identified in some fixed number of the top-ranked nearest neighbours, when averaged over a set of searches for bioactive target structures. An early example of the use of this approach is a comparison of 3D similarity measures based on inter-atomic distances ***(17)***, with Briem and Lessel providing a more recent application in their extended comparison of virtual screening methods ***(18).*** The use of a fixed cut-off means that this measure is basically a reformulation of precision, which is entirely acceptable in the early stages of a discovery programme, when the immediate need is to identify additional active molecules; however, the measure takes no account of recall, which may be an important factor in a detailed comparative study of the behaviour of different similarity measures. A second, and alternative, 'leave-one-out' classification approach assumes that the activity of one of the molecules in the database, *X*, is unknown. A similarity search is carried out using *X* as the target structure and the top-*x* (where *x* is odd) nearest neighbours identified. The activity or inactivity of *X* is then predicted on the basis of a majority vote (hence the requirement for an odd number) of the known activities of the selected nearest neighbours. This process is repeated for each of the *N* molecules in turn (or just the *A* active molecules in many cases), yielding a contingency table of the sort shown in Table 2. Various statistics can be produced from the elements of this table: perhaps the most common is Cohen's kappa statistic ***(19)***. This is defined to be

$$\frac{O-E}{1-E},$$

where *O* and *E* are the observed and expected accuracies of classification. These accuracies can be defined in terms of the elements of Table 2 as follows:

$$O = \frac{i+l}{n}, \text{ and}$$

$$E = \frac{(i+k)(i+j)+(j+l)(k+l)}{n^2}.$$

There are many variants on this basic idea, such as the weighted kappa described by Cohen himself ***(20)*** and the Rand statistic ***(21)***, which is perhaps the most widely used of the measures available for comparing different clusterings of the same set of objects.

Finally, it may be of interest to study the performance of a measure across the entire ranking resulting from a similarity search, rather than the performance for some fixed number of nearest neighbours. In this case, the most popular approach is the use of a *cumulative recall* graph, which plots the recall against the number of compounds retrieved (i.e., *a*/*A* against *n* using the notation of Table 1). The best-possible such graph would hence be one in which the *A* relevant documents are at the top of the ranking, *i.e.*, at rank-positions 1, 2, 3…*A* (or at rank-positions, *N*-*A*+1, *N*-*A*+2, *N*-*A*+3…*N* in the case of the worst-possible ranking). The use of such diagrams is exemplified by studies of similarity searching using physicochemical descriptors ***(12)*** and of a range of virtual screening methods for searching agrochemical datasets ***(22).*** The cumulative recall plot is closely related to the *receiver operating characteristic* (ROC) curves that are widely used in signal detection and classification problems ***(23)***. An ROC curve plots the true positives against the false positives for different classifications of the same set of objects; this corresponds to plotting *a* against *n-a* using the notation of Table 1, and thus the shape of an ROC curve tends to the shape of a cumulative recall plot when *n*>>*a*. An example of the use of ROC plots in chemoinformatics is provided by the work of Cuissart *et al.* on similarity-based methods for the prediction of biodegradability ***(24)***.

2.1.3. Use of quantitative data

The similar property principle can also be applied to the analysis of datasets for which quantitative bioactivity data are available, most commonly using a simple modification of the 'leave-one-out' classification approach described above. Here, the predicted property value for the target structure *X*, *P(X)*, is taken to be the arithmetic mean of the observed property values of the selected nearest neighbours. This procedure results in the calculation of a *P(X)* value for each of the *N* structures in a dataset, and an overall figure of merit is then obtained by calculating the product moment correlation coefficient between the sets of N observed and N predicted values. This approach can equally well be applied to the evaluation of clustering methods, with the predicted values here being the mean of the other compounds in the cluster containing the chosen molecule, X.

This application of the similar property principle was pioneered by Adamson and Bush ***(25, 26)*** and has since been very extensively applied. For example, Willett and Winterman used it in one of the first detailed comparisons of measures for similarity searching ***(27)*** and it also formed the basis for Brown and Martin's much-cited comparison of clustering methods and structural descriptors for compound selection ***(28)***.

2.2. Molecular Diversity Methods

2.2.1. Introduction

The principal aim of molecular diversity analysis is to identify structurally diverse (synonyms are dissimilar, disparate and heterogeneous) sets of compounds that can then be tested for bioactivity, the assumption being that a structurally diverse set will generate more structure-activity information than will a set of compounds identified at random. The sets of compounds can be selected from an existing corporate or public database, or can be the result of a systematic combinatorial library design process ***(4, 5)***.

Many of the comments that were made in Section 3.1.1 regarding similarity measures are equally applicable to diversity methods, in that the latter involve knowledge of the degree of dissimilarity or distance between pairs, or larger groups, of molecules. Here, however, there is also the need to specify a *selection algorithm*, which uses the computed dissimilarities to identify the final structurally diverse set of compounds, and there may also be a *diversity index*, which quantifies the degree of diversity in this set. It is thus important, as with similarity measures, to isolate the effect of the various components of the diversity methods that are being analysed in a comparative study. There have been many such comparisons, e.g., ***(28-33)***. Here, we focus on diversity indices since it is these that measure the overall effectiveness of a method (in fact, while an index is computed once a selection algorithm has completed its task, there are some types of algorithm that seek explicitly to optimise the chosen index, so that the current value of the index drives the operation of the selection algorithm).

Many of the early evaluations of the effectiveness of diversity methods used structure-based diversity indices, such as functions of inter-molecular dissimilarities in the context of distance-based selection methods or of the numbers of occupied cells in partition-based selection methods ***(4)***. A wide range of such indices has been reported, as discussed in the excellent review by Waldman *et al*. ***(34).*** They do, however, have the limitation that they quantify diversity in *chemical space*, whereas the principal rationale for molecular diversity methods is to maximise diversity in *biological space* ***(35)***, and we hence focus here on indices that take account of biological activity.

2.2.2. General screening programmes

We have noted the importance of the similar property principle, which would imply that a set of compounds exhibiting some degree of structural redundancy, *i.e.*, containing molecules that are near neighbours of each other, will also exhibit some degree of biological redundancy; a structurally diverse subset, conversely, should maximise the number of types of activity exhibited by its constituent molecules. It should thus be possible to compare the effectiveness of different structure-based selection methods by the extent to which they result in subsets that exhibit as many as possible of the types of activity present in the parent dataset. Maximising biological diversity in this way is the principal aim of general screening programs, which aim to select molecules from a database (or design combinatorial libraries for synthesis) that exhibit the widest possible range of different types of activity. An obvious measure of the diversity of the resulting compounds is hence the number of types of activity exhibited by them. This can be easily tested using one of the public databases that contain both chemical structures and pharmacological activities, such as the *MACCS Drug Data Report* (MDDR, at URL http://www.mdli.com/products/mddr.html) or the *World Drugs Index* (WDI, at URL http://www.derwent.com/worlddrugindex/index.html) databases. Thus, in one of the earliest comparative studies of methods for comparing diverse database subsets, Snarey *et al*. compared a range of maximum dissimilarity and sphere exclusion methods for dissimilarity-based compound selection by means of the number of different types of activity present in subsets chosen from molecules in the WDI database ***(31)***; this approach has been adopted in several subsequent studies.

2.2.3. Focused screening programmes

In a focused screening programme, the aim is to select molecules from a database (or design combinatorial libraries for synthesis) that provide the maximum amount of information about the relationships that exist between structural features and some specific type of biological activity. If this data is qualitative in nature, then a simple count of the active molecules present will suffice to quantify the degree of biological diversity. However, at least some account must additionally be taken of the chemical diversity that is present, to avoid a high level of diversity being ascribed to a cluster of highly similar molecules (such as “me too” or “fast follower” compounds in a drug database). An example of this approach is a comparison of binning schemes for cell-based compound selection by Bayley and Willett ***(36)*** that selected one molecule from each cell in a grid (thus ensuring that the selected molecules were structurally diverse) and then noted how many of these selected molecules were bioactive (thus quantifying the biological diversity).

Once interest has been focused on some small volume of structural space, large numbers of molecules are synthesised and tested (and often re-tested in the case of HTS data), and the results of these experiments used to develop a quantitative structure-activity relationship (QSAR). It has for long been claimed that the use of diverse sets of compounds will enable more robust QSARs to be developed than can be developed using randomly-chosen training sets. That this is in fact the case has been demonstrated recently by Golbraikh and Tropsha ***(37)***, and one can hence quantify the effectiveness of a compound selection method by the predictive power of the QSARs that can be derived from the compounds selected by that method. Quantitative bioactivity data also lies at the heart of the neighbourhood behaviour approach of Patterson et al. ***(33)***, which is analogous to the similar property principle but emphasises the absolute differences in descriptor values and in bioactivity values, rather than the values themselves. Specifically the authors state that a meaningful descriptor for diversity analysis is one for which “small differences in structure do not (often) produce large differences in biology”, and then use this idea to compare a wide range of descriptor types by means of a χ^2 analysis; an improved version of this analysis is described by Dixon and Merz ***(38)***.

3. Notes

1. The group in Sheffield has over two decades experience of carrying out comparative studies of similarity (and, more recently, diversity) methods. Perhaps the most importance single piece of advice we can give to those wishing to carry out comparable studies is the need to use a range of types of data, ideally including both homogeneous and heterogeneous datasets. Only by so doing can one ensure the robustness and general applicability of the methods that are being compared. In particular, one would not wish to encourage the situation that pertained for some time in the QSAR literature, where a new method was normally developed and tested on a single dataset, most commonly the set of steroids ***(39)*** first popularised by Cramer *et al.* ***(40).***
2. In like vein, we would recommend the use of more than just one evaluation measure. That said, it is our experience that different measures usually agree as to the relative merits of different approaches (unless there are only very minor differences in effectiveness): even so, it is always worth carrying out additional analyses to ensure that one's results are, indeed, independent of the evaluation criterion that has been adopted.
3. Having criticised the exclusive use of the steroid dataset, it does have the great advantage that it provides a simple basis for comparison with previous work, and it would be highly desirable if comparable test-sets were available for similarity and diversity analyses. To some extent, this is already happening with increasing use being made of the qualitative bioactivity data in the MDDR and WDI datasets mentioned previously; two other datasets that can be used for this purpose, and which have the advantage that they are available for public download, are the cancer and AIDS files produced by the National Cancer Institute (at URL http://dtp.nci.nih.gov/).

References

1. Johnson, M. A. and Maggiora, G. M. (eds.) (1990) *Concepts and Applications of Molecular Similarity*. Wiley, New York.
2. Dean, P. M. (ed.) (1994) *Molecular Similarity in Drug Design*. Chapman and Hall, Glasgow.

3. Willett, P., Barnard, J. M. and Downs, G. M. (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**, 983--996.

4. Dean, P. M. and Lewis, R. A. (eds.) (1999) *Molecular Diversity in Drug Design*. Kluwer, Amsterdam.

5. Ghose, A. K. and Viswanadhan, V. N. (eds.) (2001) *Combinatorial Library Design and Evaluation: Principles, Software Tools and Applications in Drug Discovery*. Marcel Dekker, New York.

6. Kubinyi, H. (1998) Similarity and dissimilarity – a medicinal chemist's view. *Perspect. Drug. Discov. Design* **11**, 225--252.

7. Martin, Y. C., Kofron, J. L. and Traphagen, L. M. (2002) Do structurally similar molecules have similar biological activities? *J. Med. Chem.* **45**, 4350--4358.

8. Salton, G. and McGill, M. J. (1983) *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

9. Frakes, W.B. and Baeza-Yates, R. (eds.) (1992) *Information Retrieval: Data Structures and Algorithms*. Prentice Hall, Englewood Cliffs NJ.

10. Sparck Jones, K. and Willett, P. (eds.) (1997) *Readings in Information Retrieval*. Morgan Kaufmann, San Francisco CA

11. Edgar, S. J., Holliday, J. D. and Willett, P. (2000) Effectiveness of retrieval in similarity searches of chemical databases: a review of performance measures. *J. Mol. Graph. Model.* **18**, 343--357.

12. Kearsley, S. K., Sallamack, S., Fluder, E. M., Andose, J. D., Mosley, R. T. and Sheridan, R. P. (1996) Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **36**, 118--127.

13. Gillet, V. J., Willett, P. and Bradshaw, J. (1998) Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* **38**, 165--179.

14. Güner, O.F. and Henry, D.R. Formula for determining the "goodness of hit lists" in 3D database searches. At URL http://www.netsci.org/Science/Cheminform/feature09.html.

15. Güner, O.F. and Henry, D.R. (2000) Metric for analyzing hit lists and pharmacophores, in *Pharmacophore Perception, Development and Use in Drug Design* (Guner, O., ed.), International University Line, La Jolla CA. pp. 193--212

16. Raymond, J. W. and Willett, P. (2002) Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *J. Comput.-Aid. Mol. Design* **16**, 59--71.

17. Pepperrell, C. A. and Willett, P. (1991) Techniques for the calculation of three-dimensional structural similarity using inter-atomic distances. *J. Comput.-Aided Mol. Design* **5**, 455--474.

18. Briem, H. and Lessel, U. F. (2000) *In vitro* and *in silico* affinity fingerprints: finding similarities beyond structural classes. *Perspect Drug Discov. Design* **20**, 231--244.

19. Cohen, J. A. (1960) A coefficient of agreement for nominal scales. *Educ. Psychol. Measure*. **20**, 37--46.

20. Cohen, J. A. (1968) Weighted kappa: nominal scale agreement with provision fro scale disagreement or partial credit. *Psychol. Bull.* **70**, 213--220.

21. Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods. *J. Amer. Stat. Assoc*. **66**, 846--850.

22. Wilton, D., Willett, P., Mullier, G. and Lawson, K. (2003) Comparison of ranking methods for virtual screening in lead-discovery programmes. *J. Chem. Inf. Comput. Sci.* **43**, 469--474.

23. Egan, J. P. (1975) *Signal Detection Theory and ROC Analysis*, Academic Press, New York.

24. Cuissart, B., Touffet, F., Crémilleux, Bureau, R. and Rault, S. (2002) The maximum common substructure as a molecular depiction in a supervised classification context: experiments in quantitative structure/biodegradability relationships. *J. Chem. Inf. Comput. Sci*. **42**, 1043--1052.

25. Adamson, G. W. and Bush, J. A. (1973) A method for the automatic classification of chemical structures. *Inf. Stor. Retriev.* **9**, 561--568.

26. Adamson, G. W. and Bush, J. A. (1975) A comparison of the performance of some similarity and dissimilarity measures in the automatic classification of chemical structures. *J. Chem. Inf. Comput. Sci.* **15**, 55--58.

27. Willett, P. and Winterman, V. (1986) A comparison of some measures for the determination of inter-molecular structural similarity. *Quant. Struct.-Activ. Relat.* **5**, 18--25.

28. Brown, R. D. and Martin, Y. C. (1996) 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.

29. Brown, R. D. (1997) Descriptors for diversity analysis. *Perspect. Drug Disc. Design* **7/8**, 31--49.

30. Bayada, D. M., Hamersma, H. and van Geerestein, V. J. (1999) Molecular diversity and representativity in chemical databases. *J. Chem. Inf. Comput. Sci*. **39**, 1--10.

31. Snarey, M., Terret, N. K., Willett, P. and Wilton, D. J. (1997) Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graph. Model.* **15**, 372--385.

32. Matter, H. and Potter, T. (1999) Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J. Chem. Inf. Comput. Sci*. **39**, 1211--1225.

33. Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D. and Weinberger, L. E. (1996) Neighbourhood behaviour: a useful concept for validation of “molecular diversity” descriptors. *J. Med. Chem.* **39**, 3049--3059.

34. Waldman, M., Li, H. and Hassan, M. (2000) Novel algorithms for the optimisation of molecular diversity of combinatorial libraries. *J. Mol. Graph. Model*. **18**, 412--426.

35. Ferguson, A. M., Patterson, D. E., Garr, C. D. and Underiner, T. L. (1996) Designing chemical libraries for lead discovery. *J. Biomol. Screen.* **1**, 65--73.

36. Bayley, M. J. and Willett, P. (1999) Binning schemes for partition-based compound selection. *J. Mol. Graph. Model.* **17**, 10--18.

37. Golbraikh, A. and Tropsha, A. (2002) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput.-Aid. Mol. Design* **16**, 357--369.

38. Dixon, S. L. and Merz, K. M. (2001) One-dimensional molecular representations and similarity calculations: methodology and validation. *J. Med. Chem.* **44**, 3795—3809.

39. Coats, E. A. (1998) The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods. *Perspect. Drug Discov. Design* **12/14**, 199--213.

40. Cramer, R. D., Patterson, D. E. and Bunce, J. D. (1988) Comparative Molecular Field Analysis (CoMFA). Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc*. **110**, 5959--5967.

Table 1. Contingency table describing the output of a search in terms of active molecules and molecules retrieved in a similarity search retrieving n molecules.

		Active		
		Yes	No	
Retrieved	Yes	a	n-a	n
	No	A-a	N-n-A+a	N-n
		A	N-A	N

Table 2. Contingency table describing the output of a search in terms of correctly and incorrectly predicted molecules in a classification experiment classifying n molecules.

		Classification		
		Active	Inactive	
Truth	Active	i	j	i+j
	Inactive	k	l	k+l
		i+k	j+l	n