Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals

Robin Taylor†

Zeneca Agrochemicals, Jealott's Hill Research Station, Bracknell, Berkshire RG12 6EY, UK

Received June 16, 1994⊗

Computer simulations have been performed to investigate the effectiveness of experimental design techniques when applied to pharmaceutical and agrochemical random screening. Two design algorithms have been investigated, a maximum dissimilarity technique (stepwise elimination) and an approach based on sampling clusters. Results show that the former technique tends to *reduce* the chances of finding active molecules quickly. In contrast, the cluster sampling approach can afford small improvements in the rate at which active molecules are identified. However, use of maximum dissimilarity methods may still be beneficial if a premium is placed on finding active molecules with unusual structures.

INTRODUCTION

Random screening has been, and remains, one of the most successful methods of lead generation in the drug and agrochemical industries. It involves testing large numbers of random chemical compounds on a biological assay in the hope that compounds with useful effects will be detected. The assays may be in vivo or in vitro. The compounds are usually selected from in-house collections, of the sort assembled by most chemical companies over the course of time.

A key feature of random screening is that the experimentalist has few or no preconceptions about the molecular features required for activity. For this reason, compounds are usually tested in an arbitrary order. In recent years, however, it has been suggested that experimental design techniques may be of use in selecting the order in which compounds are screened. Two strategies have been proposed: maximum dissimilarity selection[1] and cluster sampling.[2] The former is used to arrange the available compounds into a "rational" order, such that those screened first are as structurally diverse as possible. The latter is used to locate clusters of close structural analogues, from each of which can be selected one or two representative compounds. The techniques differ in that maximum dissimilarity methods tend to pick preferentially a high proportion of "outliers" (i.e., molecules that have no close structural analogues in the set), whereas cluster sampling does not. For example, consider Figure 1, which shows a map of a hypothetical molecular-structure space occupied by 15 molecules. If a maximum dissimilarity method were used to select four of the molecules for screening, it might typically pick 1, 4, 9, and 15, affording good coverage of the space. A cluster sampling algorithm would focus on densely occupied regions of the space and hence avoid outliers such as 1, 9, and 15; a typical cluster-sampling selection would be 2, 5, 7, and 11.

In principle, there could be two reasons for using experimental design techniques: to increase the *novelty* of hits (i.e., increase the chances of finding activity in molecules with unusual structures, which could be an advantage in terms of patentability) and to improve the *hit rate* (i.e., find active molecules as quickly as possible, regardless of whether or not they have unusual structures). By ensuring that outliers are screened at an early stage, maximum dissimilarity methods are likely to achieve the former aim. However, it is less clear that either design strategy will improve hit rate, and it is this issue that is addressed here.

It is assumed below that, given a new biological assay and a set of random compounds, the initial probability of activity is equal for all members of the set (this assumption is discussed later). It follows that no experimental design technique *on its own* can increase the likelihood of finding active molecules quickly. For example, maximum dissimilarity methods can be used to order the entire set of compounds so that structurally diverse molecules are tested as soon as possible: but it remains the case that an active molecule is, at this stage, just as likely to be in the bottom half of the list as in the top half.

The point of using design techniques is to maximize the information gained about structure−activity relationships in the early stages of screening. For example, both maximum dissimilarity and cluster sampling techniques will ensure that data are generated quickly about the variation of activity across a wide range of structural types. However, in order to find active molecules quickly, this information must be *used*. There are two ways in which this can be done. Firstly, if a molecule is observed to be active—even if only weakly—then close structural analogues may be promoted in the order (i.e., tested earlier). This is *positive feedback*. Secondly, if a molecule is observed to be inactive, similar molecules can be relegated down the order—*negative feedback*.

This paper describes computer simulations that estimate the relative effects on hit rate of experimental design techniques and positive and negative feedback. Other factors that may be relevant are also investigated, such as screen precision and the time required for testing a compound. The paper begins with a description of the algorithms used and of the simulation model.

† Current address: Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK.

⊗ Abstract published in *Advance ACS Abstracts,* November 15, 1994.

0095-2338/95/1635-0059$09.00/0

Table 1. Brief Description of Simulation Parameters

P_BACKGROUND	probability of a compound being weakly active if it is not structurally similar to the highly active molecule
P_RATIO	defines probability of weak activity in a compound that is structurally similar to the highly active one; this probability is equal to the product of P_RATIO and P_BACKGROUND
SIM_THRESHOLD	defines whether a molecule is considered similar to the highly active molecule (deemed similar if the similarity coefficient of the two molecules exceeds SIM_THRESHOLD)
P_FUNCTION	defines probability function for assignment of weak activity (STEP or LINEAR, see Figure 2)
P_FALSE_MISS	probability of a weakly active molecule being falsely "detected" as inactive
P_FALSE_HIT	probability of an inactive molecule being falsely "detected" as weakly active
N_BATCH	number of compounds in a batch
PRE_DELAY	defines time required for assembly of a batch: the mth batch must be chosen immediately before the (m − *PRE_DELAY*)th is tested
POST_DELAY	defines time required for processing results of a batch: results for mth batch become available after the (m + *POST_DELAY*)th is tested
N_POSFEED	defines extent to which positive feedback used; if a weakly active molecule is detected, N_POSFEED of its near neighbors will be screened as soon as possible
POS_STRAT	defines whether positive feedback is used recursively
P_NEGFEED	defines extent to which negative feedback used: the closer P_NEGFEED is to one, the more stringent the negative feedback

COMPUTER ALGORITHMS

All programs were written in Fortran 77 on a Silicon Graphics Power Iris 4D/220, running under Irix.

Estimation of Molecular Similarity. Use of experimental design techniques requires a method for estimating the similarity of any given pair of molecules.[3] In this work, each molecule was characterized by a set of substructure keys, each key being set to one or zero depending on the presence or absence, respectively, of a particular substructure in the molecule. The keys used were the user-searchable set provided in the MACCS software system.[4] The similarity of molecules i and j was then estimated by the weighted Tanimoto coefficient:[5]

$$s_{ij} = \frac{\sum_m w_m k_{im} k_{jm}}{\sum_m w_m k_{im} + \sum_m w_m k_{jm} - \sum_m w_m k_{im} k_{jm}}$$

Here, k_{im} and k_{jm} are the values of the mth key in molecules i and j, respectively, and w_m is a weight reflecting the importance of the substructural feature corresponding to the mth key. For example, substructures containing hydrogen-bonding atoms were given larger weights than those containing only carbon atoms, since hydrogen bonds are usually an important element of molecular recognition.

The results described below depend to some extent on the choice of keys, weights, and similarity coefficient. However, as it turns out, the key conclusions of the study are independent of these factors.

Maximum Dissimilarity Selection: Stepwise Elimination. A variety of maximum dissimilarity algorithms can be envisaged. In this work, an in-house stepwise elimination (SE) procedure was used. The procedure begins with the calculation of the similarity matrix, **S**, which is a symmetric matrix of order n (n = number of available compounds), the ijth element of which is the similarity coefficient of molecules i and j, s_{ij}. The matrix is scanned to find its largest element, which of course corresponds to the most similar pair of molecules in the set. One of this pair is selected at random and eliminated, producing a reduced similarity matrix of order $n - 1$. Repetition of this scanning/elimination process successively reduces the number of molecules until only one remains. As the algorithm proceeds, it converges on a subset of very dissimilar molecules. The final step of the procedure is therefore to arrange the compounds in the opposite order to which they were eliminated. This places structurally diverse molecules at the top of the list.

Cluster Sampling. Cluster sampling (CS) was performed by an algorithm based on analysis of the nearest neighbor table (NNT). For a set of n compounds, the NNT contains n rows, one per compound. The ith row is simply a list of the molecules in the set which are similar to compound i (its "near neighbors"). Molecules were deemed similar if their similarity coefficient exceeded 0.8. In the event of a compound having a very large number of near neighbors, only the 300 most similar were stored.

The CS algorithm then proceeds as follows. The first molecule to be selected is the one that occurs most often in the NNT, since this will tend to be in the center of the most densely occupied region of the molecular-structure space. All near neighbors of the selected molecule are then *held*, which means that they are made unavailable for selection. The next molecule to be picked is the one amongst those not held that occurs most often in the NNT. It will tend to be in the center of the second most densely occupied region of molecular-structure space. Its near neighbors are held. The selection/holding process is repeated until all molecules have either been chosen or held. At this point, the latter are released (made available for selection) and the selection/holding process resumed. After sufficient cycles, all molecules will have been selected.

Unlike conventional cluster analysis algorithms, such as the Jarvis−Patrick method,[2,6,7] CS never explicitly partitions the set of molecules into clusters. However, it will tend to order the molecules so that the natural clusters in the set, even though undetermined, are sampled systematically, starting with the largest and moving toward the smallest (i.e., the outliers).

SIMULATION PROGRAM

Overview. The usual result of a random screening exercise is that the majority of compounds are inactive, a few show weak activity, and only very occasionally is a highly active molecule found. The weak activities are worth knowing about, but interest mainly centers on the compounds

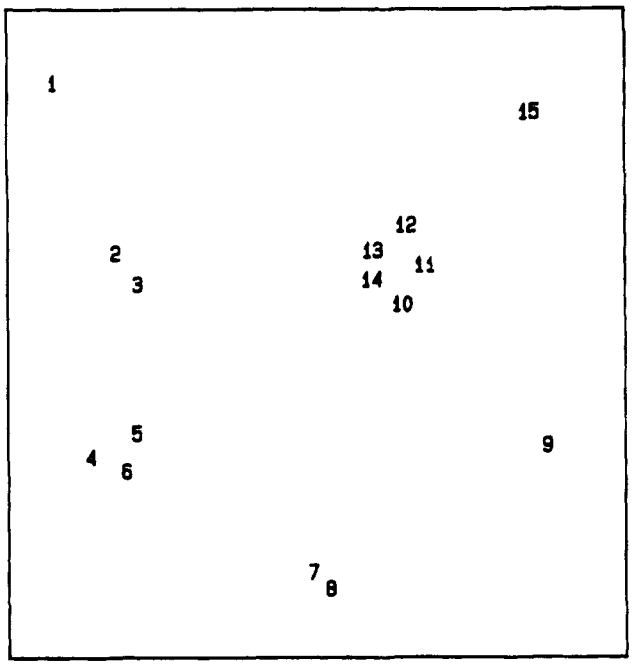

Figure 1. Map showing 15 molecules distributed in a hypothetical molecular-structure space.

with high activity. The simulation program was therefore designed to investigate how quickly such molecules can be found.

The program took as input the structures of a set of random molecules. Each simulation consisted of a number of cycles, usually between 1000 and 20 000. The procedure in each cycle was as follows. Firstly, a random number generator was used to assign activities to the molecules, so that one molecule was very active, a few were weakly active, and the rest (being the majority) were inactive. The compounds were then arranged into an initial order for testing, either at random or by using the SE or CS algorithms. The screening process was then simulated, the compounds being tested in batches. As biological results were generated, the program revised the order in which the remaining compounds were to be screened by using positive and negative feedback. A given cycle stopped when the highly active molecule was found. The program stored the number of compounds that had to be tested to find the highly active molecule (= n).

After all cycles were completed, the average of the n values (= N) was computed. This was the primary result of the simulation, and its variation with design strategy, etc., the major focus of interest.

The following paragraphs describe the details of the simulation procedure and define the model parameters. Table 1 summarizes these parameters.

Assignment of Biological Activities. Before each cycle of simulation, biological activities were assigned to all molecules by random number generation. First, one molecule was picked arbitrarily and assigned high activity; all molecules had an equal probability of being chosen. Once this selection was made, a few of the remaining molecules were assigned weak activity; the rest were inactive. Importantly, the probability of a given molecule being assigned weak activity depended on how similar it was to the highly active compound. Two probability functions were tried, a step function (P_FUNCTION = STEP) and a linear function (P_FUNCTION= LINEAR). With the STEP function, the

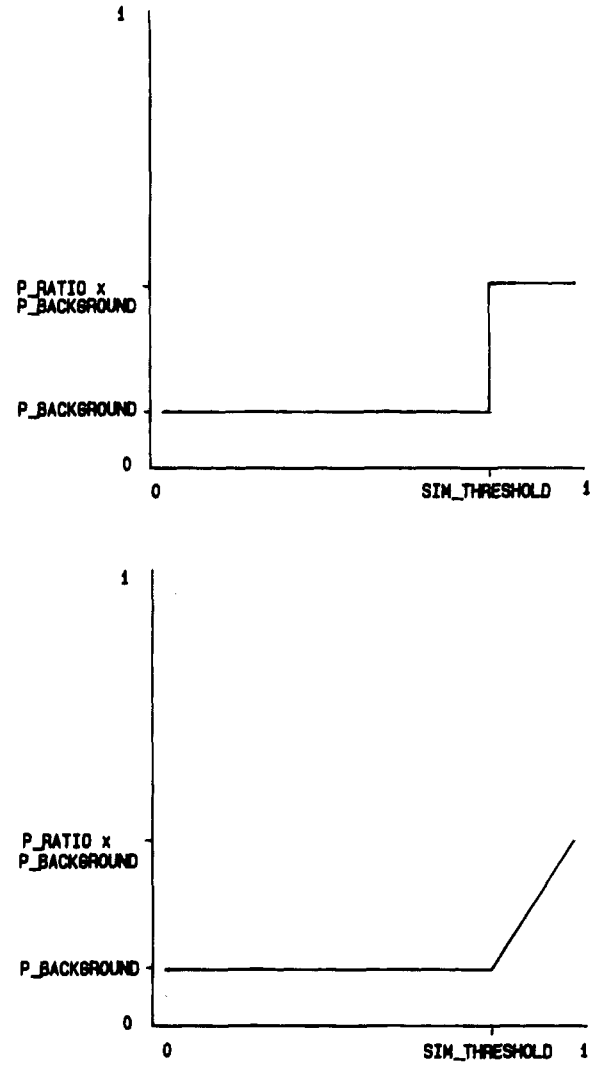

Figure 2. Probability distributions used for assignment of weak activity in simulation, when P_FUNCTION is set at (a, top) STEP and (b, bottom) LINEAR. Vertical axis plots probability, horizontal axis plots similarity coefficient with highly active molecule.

probability of the ith molecule being weakly active varied with s_{ia}—its similarity coefficient with the very active molecule—as shown in Figure 2a. In this figure, P_BACKGROUND, P_RATIO, and SIM_THRESHOLD are user-defined parameters. P_BACKGROUND is the *background probability* of weak activity, i.e., the probability amongst molecules dissimilar to the very active one ($s_{ia} \leq$ SIM_THRESHOLD). Compounds similar to the highly active molecule ($s_{ia} >$ SIM_THRESHOLD had an elevated probability of weak activity, equal to the product of P_RATIO and P_BACKGROUND, where P_RATIO exceeds unity. This parameter is effectively a measure of the strength of structure−activity correlation in the neighborhood of the highly active molecule.

The alternative LINEAR function is shown in Figure 2b. Again, compounds similar to the highly active molecule have an increased probability of weak activity. This varies linearly from a maximum of P_RATIO.P_BACKGROUND at s_{ia} = 1 to P_BACKGROUND at s_{ia} = SIM_THRESHOLD.

Screen Precision. Experimental errors were built into the model by introducing some uncertainty into the "measurement" of activity. Thus, during the simulation, a weakly active molecule could be falsely recorded as inactive, with a probability of P_FALSE_MISS. The probability of the

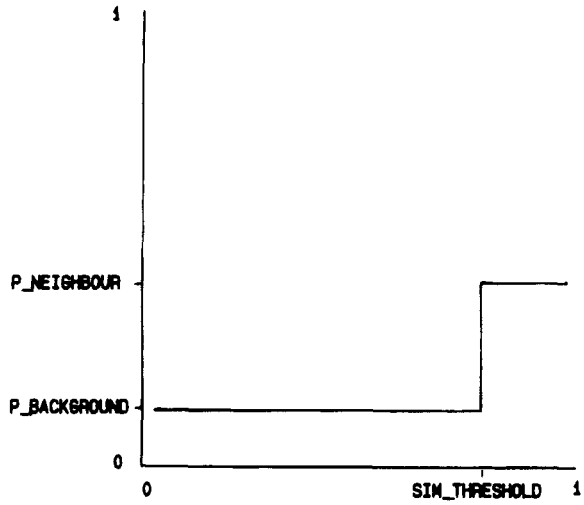

Figure 3. Alternative probability distribution for assignment of weak activities. Axes definitions as in Figure 2.

opposite error (inactive molecule recorded as weakly active) was P_FALSE_HIT.

Assembly of Compound Batches. Compounds were tested in batches of size N_BATCH. Pre- and post-processing delays were built into the model, to simulate the time required in a real experiment for sample preparation, measurement, and analysis of the biological activities. Thus, during the simulation, the mth batch had to be selected just before the (m − PRE_DELAY)th was tested, where *PRE_DELAY* is a user defined integer parameter. Results for the mth batch became available (and hence could be used for positive and negative feedback) after the (m +*POST*_DELAY)th batch was tested.

Positive Feedback. Under most circumstances (but see below), detection of weak activity in a compound resulted in the program selecting a certain number of its near neighbors and placing them in the next available batch. This action was taken whether the weakly active molecule was a genuine or a false hit (see Screen Precision, above). "Near neighbors" were molecules whose similarity coefficients with the weakly active molecule exceeded SIM_THRESHOLD, a user-defined parameter which was set at 0.8 in most of the simulations. The number of near neighbors selected was N_POSFEED, another user-defined parameter. If there were more than N_POSFEED untested near neighbors available, those with the largest similarity coefficients were selected. If there were fewer than N_POSFEED, then all were taken.

Two positive feedback strategies were tried (POS_STRAT = SINGLE or RECURSIVE). The difference between them is best explained by an example. Suppose that the very first molecule to be tested was weakly active. N_POSFEED of its near neighbors would be placed in the next available batch. Suppose that one of these was also weakly active. If POS_STRAT = RECURSIVE, then N_POSFEED near neighbors of this compound would also be selected for immediate testing. However, if POS_STRAT = SINGLE, the positive feedback would be ended after the first set of near neighbors was tested, irrespective of their activities.

Negative Feedback. Before a compound was accepted for a batch, the results of any tests already performed on its near neighbors (compounds with similarity coefficients > SIM_THRESHOLD) were reviewed. Suppose there were n such tested near neighbors, of which m showed weak activity (either genuine or because they were false hits). The compound was not accepted for the batch, but instead put to the bottom of the untested compound list, if $b <$ P_NEGFEED. Here, P_NEGFEED is a user-specified parameter, and b is the binomial probability of there being m or fewer than m successes in n trials when the probability of an individual success is P_BACKGROUND. The closer P_NEGFEED is to unity, the more stringent is the negative feedback, i.e., the more likely it is that a compound will be rejected because too many of its tested near neighbors are inactive.

Test Sets of Molecules. Two sets of molecules were used, each comprising 1000 members. Both were subsets of the Zeneca Agrochemicals compound database. Set 1 consisted of random molecules registered into the database during a single period of a few months in the mid 1980s. Set 2 contained about 100 molecules from each of several years in that decade. Probably as a result, set 2 is more structurally diverse than set 1 (i.e., contains fewer large clusters).

RESULTS

In the discussion below, N is the average number of compounds that had to be tested in a simulation before the highly active molecule was found, the average being taken over all cycles performed. Since there were 1000 molecules in each test set, the random expectation value of N is 500.5.

Convergence. An initial series of simulations was performed on set 1 to determine the convergence properties of the program. Twenty-two representative sets of model parameters were chosen and simulations performed for each. In each cycle, the compounds were arranged initially into a random order, but this was of course refined by positive and negative feedback as the cycle proceeded. Results were printed out after 1000, 2500, 5000, 10 000, and 20 000 cycles. Table 2a summarizes the results and gives the rms deviations between the intermediate and final N values. Results at 1000 cycles deviate appreciably from the final (20 000 cycle) values, but reasonable convergence has been achieved at 2500 cycles.

As a further test, a simulation with one particular set of model parameters (Table 2b) was repeated 63 times, using set 1 and random initial ordering. The standard deviations of the N values obtained after 20 000 cycles was 1.52, compared with 5.64 after 2500 cycles and 8.28 after 1000 cycles. These standard deviations are used below to estimate the significance of differences between various N values in the main simulations.

Test of Random Expectation. Sixty-three simulations (each of 2500 cycles) were performed on set 1, using random initial ordering and setting P_RATIO = 1. The average of the N values obtained was 500.7, with a standard error of 0.8. This is not significantly different from the random expectation value of 500.5. The result is to be expected, since if P_RATIO = 1, there is no structure−activity correlation in the neighborhood of the highly active molecule; in other words, its near neighbors are no more likely to show weak activity than any other molecule in the set. In consequence, there is no information which can guide the simulated screening experiment toward the highly active molecule. It is evident that the CS and SE algorithms cannot help in this situation. Thus, it comes as no surprise that the average values of N from 63 simulations with P_RATIO = 1 were 501.2 when the initial order was chosen by CS and 500.6 when it was chosen by SE.

Table 2. Convergence Properties of Simulation Program

(a) Values of N After Various Numbers of Cycles in 22 Different Simulations

simulation no.	number of cycles				
	1000	2500	5000	10 000	20 000
1	432	422	426	424	424
2	305	319	320	318	322
3	361	381	384	379	375
4	376	381	381	377	381
5	447	446	444	451	448
6	443	438	437	437	439
7	303	307	305	311	306
8	370	375	371	366	368
9	372	373	378	374	377
10	452	445	437	442	445
11	402	400	410	402	406
12	399	409	400	406	403
13	448	455	459	460	457
14	457	450	454	458	450
15	476	473	465	472	466
16	477	464	465	478	471
17	486	473	470	474	475
18	479	473	472	472	467
19	398	397	402	401	398
20	485	489	490	490	489
21	512	486	490	499	498
22	483	490	497	494	496
RMS[a]	8.6	4.8	4.3	3.9	0.0

(b) Parameter Values Used for Estimating Standard Deviation of N

P_BACKGROUND	0.10	N_BATCH	1
P_RATIO	7.0	PRE_DELAY	1
SIM_THRESHOLD	0.80	POST_DELAY	0
P_FUNCTION	STEP	N_POSFEED	5
P_FALSE_MISS	0.30	POS_STRAT	RECURSIVE
P_FALSE_HIT	0.10	P_NEGFEED	0.90

[a] RMS = rms deviation between intermediate and final N values.

Table 3. Parameter Values Used in Sensitivity Analysis

P_BACKGROUND	0.01, 0.05, 0.10	N_BATCH	1, 10, 20
P_RATIO	1.2, 2.0, 5.0	PRE_DELAY	1, 2, 4
SIM_THRESHOLD	0.79, 0.80[a]	POST_DELAY	0, 2, 4
P_FUNCTION	STEP, LINEAR	N_POSFEED	1, 5, 10
P_FALSE_MISS	0.0, 0.1, 0.3	POS_STRAT	SINGLE, RECURSIVE
P_FALSE_HIT	0.0, 0.1, 0.3	P_NEGFEED	0.80, 0.95, 0.99

[a] Several pairs of molecules in the set had similarity coefficients of exactly 0.8, so the apparently small change in SIM_THRESHOLD from 0.79 to 0.80 had a disproportionately large effect on the number of pairs of molecules that were considered near neighbors in the simulations.

Sensitivity Analysis. A series of simulations was performed to identify which parameters are most important in determining hit rate. An initial guess was made at the ranges that the various parameters might reasonably span. For some parameters these ranges were quite large, so some intermediate values were chosen (Table 3). An exhaustive investigation of all combinations of these parameter values was impracticable, there being over 150 000 in total. A random number generator was therefore used to select 3000 of the combinations arbitrarily, and a simulation was performed for each. Set 1 was used as input, with initial random ordering of the molecules. (Parallel series of simulations with the initial order chosen by CS or SE produced essentially identical conclusions and are therefore not described here.) Each simulation was run for 1000 cycles, which gives only mediocre convergence (see above) but was considered sufficient for this preliminary analysis.

The results are summarized in Table 4, which shows how the average value of N varies with each model parameter in turn. For example, the first line of the table shows that P_BACKGROUND was set to 0.01 in 1026 of the simulations, amongst which the average value of N was 496.7. The standard error of this mean can be estimated as approximately $8.28/\sqrt{1026} = 0.3$, since the standard deviation of N in a single simulation of 1000 cycles is about 8.28 (see above). Standard errors of all other mean values in Table 4 were estimated in similar fashion.

The table suggests that N is most dependent on P_RATIO, P_BACKGROUND, P_FALSE_HIT, and P_FUNCTION. It also varies to a smaller extent with four other parameters, P_FALSE_MISS, N_POSFEED, P_NEGFEED, and SIM_THRESHOLD. However, the remaining four parameters may be rejected as unimportant. Thus, N shows no dependence at all on POS_STRAT, and only very minor dependence on the three batch assembly parameters. (It is worth noting, however, that the small variation of N with the latter is in the expected direction. Thus, N rises as N_BATCH, PRE_DELAY, and POST_DELAY are increased. In each case, the effect of increasing the model parameter is to introduce delays into the operation of positive and negative feedback, which may be expected to reduce feedback efficiency.)

Of the eight significant parameters, three are related, viz., P_RATIO, P_FUNCTION, and SIM_THRESHOLD. Each of them defines, to some extent, the strength of the structure–activity correlation in the neighborhood of the highly active molecule. There seemed little point in allowing all three to vary independently, so it was decided to keep P_FUNCTION and SIM_THRESHOLD fixed in future simulations. This left six parameters to vary: P_BACKGROUND, P_RATIO, P_FALSE_MISS, P_FALSE_HIT, N_POSFEED, and P_NEGFEED.

Dependence of Hit Rate on Model Parameters: Comparison of Design Techniques. Table 5 (first two columns and footnote) lists the parameter values used in the next series of simulations. All possible combinations of these values were used, there being only 486 in total. Six 2500-cycle simulations were performed for each combination, using set 1 and then set 2, with the initial order of testing chosen at random, by CS and by SE. Results are given in Table 5, columns a–f. Standard errors are about 0.4 for all the average N values given; this is based on the assumption that the standard deviation of N from a single 2500-cycle simulation is 5.64 (see above).

The results in Table 5 (set 1, random initial ordering) show that N is most dependent on P_RATIO and P_BACKGROUND. The lowest value of N is obtained when the product of these parameters is highest. The explanation is straightforward: at this point, the probability is maximized that near neighbors of the highly active molecule will themselves be weakly active. The more of these weak actives there are, the better the chance that one of them will be detected quickly and the highly active molecule found in the ensuing positive feedback. Interestingly, N remains close to its random expectation value of 500.5 when P_RATIO = 2, and substantial improvements (savings of > 50 tests) are made only when P_RATIO is as high as 10. The dependence

Table 4. Sensitivity Analysis[a]

parameter	value	NSIM	N(av)	parameter	value	NSIM	N(av)
P_BACKGROUND	0.01	1026	496.7(3)	N_BATCH	1	978	480.6(3)
	0.05	997	481.1(3)		10	1048	482.2(3)
	0.10	977	466.9(3)		20	974	482.7(3)
P_RATIO	1.2	983	498.0(3)	PRE_DELAY	1	1034	481.8(3)
	2.0	1014	488.4(3)		2	984	481.0(3)
	5.0	1003	459.3(3)		4	982	482.7(3)
SIM_THRESHOLD	0.79	1498	479.8(2)	POST_DELAY	0	978	481.0(3)
	0.80	1502	483.9(2)		2	961	482.1(3)
					4	1061	482.3(3)
P_FUNCTION	STEP	1490	476.7(2)	N_POSFEED	1	1017	484.7(3)
	LINEAR	1510	486.9(2)		5	977	482.2(3)
					10	1006	478.6(3)
P_FALSE_MISS	0.0	1006	478.6(3)	POS_STRAT	SINGLE	1510	481.9(2)
	0.1	988	481.7(3)		RECURSIVE	1490	481.8(2)
	0.3	1006	485.1(3)				
P_FALSE_HIT	0.0	1020	476.9(3)	P_NEGFEED	0.8	976	484.8(3)
	0.1	1001	480.2(3)		0.95	1032	480.2(3)
	0.3	979	488.6(3)		0.99	992	480.6(3)

[a] Standard errors in parentheses; NSIM = number of simulations performed with indicated parameter value; N(av) = average number of compounds tested before highly active molecule found.

Table 5. Comparison of Experimental Design Strategies

		N(av)					
parameter	value	*a*	*b*	*c*	*d*	*e*	*f*
P_BACKGROUND	0.005	492.4	492.9	494.5	500.1	498.0	499.2
	0.01	484.6	484.9	487.9	498.0	495.6	497.6
	0.05	435.3	433.0	452.8	485.1	476.2	491.2
P_RATIO	2.0	493.2	493.6	494.8	499.4	497.7	499.3
	5.0	473.2	472.8	480.0	495.3	491.4	497.2
	10.0	445.9	444.3	460.3	488.5	480.6	491.5
P_FALSE_MISS	0.0	466.6	465.8	475.3	493.6	487.9	495.1
	0.1	469.8	468.9	477.1	494.1	489.7	496.0
	0.3	475.9	476.0	482.7	495.5	492.1	496.9
P_FALSE_HIT	0.0	459.8	460.8	471.3	492.2	487.2	495.1
	0.1	468.8	468.6	477.3	494.3	490.1	495.2
	0.3	483.8	481.4	486.5	496.7	492.5	497.7
N_POSFEED	5	472.5	473.3	479.6	494.2	489.8	496.0
	10	469.6	469.2	478.1	494.7	489.8	495.9
	15	470.3	468.2	477.4	494.2	490.2	496.0
P_NEGFEED	0.950	472.1	471.8	479.3	495.2	490.2	496.3
	0.995	469.5	468.7	477.5	493.6	489.6	495.7

[a] Set 1, random initial order. [b] Set 1, cluster sampling. [c] Set 1, stepwise elimination. [d] Set 2, random. [e] Set 2, cluster sampling. [f] Set 2, stepwise elimination. Remaining parameters were fixed at SIM_THRESHOLD = 0.80, P_FUNCTION = STEP, N_BATCH = 10, PRE_DELAY = 1, POST_DELAY = 1, and POS_STRAT = SINGLE.

of N on P_RATIO and P_BACKGROUND is investigated in more detail later.

As expected, N increases as the screen gets less precise, i.e., as P_FALSE_HIT and P_FALSE_MISS increase. False hits have a more detrimental effect than false misses, presumably because they lead the screening experiment down "blind alleys". The dependence of N on P_NEGFEED is small but statistically significant. Values close to unity are favorable, implying that stringent negative feedback should be used. N_POSFEED has only a small effect on N in the range investigated; values toward the top end of the range appear most favorable.

Comparison of columns a, b, and c of Table 5 shows that the CS algorithm achieves little if any improvement in N, whilst the SE algorithm is actually counterproductive, i.e., decreases the probability of finding the highly active molecule quickly. This is particularly true at high values of P_RATIO. At first sight, the result appears counterintuitive, but it is statistically significant and occurred in virtually all of the simulations. The poor performance of CS and SE is discussed and rationalized in the final section of this paper.

The results obtained for set 2 (columns d−f of Table 5) differ from those for set 1 in two main respects. Firstly, the values of N are invariably higher—indeed, they are little better than the random expectation value of 500.5. Secondly, whilst SE is still counterproductive, CS now affords small but clear improvements over random ordering. The key feature in distinguishing sets 1 and 2 is that the latter is much more diverse, i.e., contains fewer large clusters of structurally similar molecules. The difference in simulation results must therefore be ascribed to this, a point which is discussed in more detail later.

Effect on Hit Rate of P_BACKGROUND. In the simulations described above, the probability of weak activity amongst near neighbors of the highly active molecule was

$$\text{P\_NEIGHBOUR} = \text{P\_RATIO.P\_BACKGROUND}$$

where P_RATIO and P_BACKGROUND are user-defined parameters (see Figure 2a). One problem with this formalism is that it is impossible to separate the effects on N of P_BACKGROUND and P_NEIGHBOUR. A final series of simulations was therefore performed with a slightly different model: a step function similar to that shown in Figure 2a was used for assignment of weak activity or inactivity, but the user-specified parameters were P_BACKGROUND and P_NEIGHBOUR, not P_BACKGROUND and P_RATIO (Figure 3).

Simulations were performed on set 1, using 20 000 cycles. Four parameters were varied: P_NEIGHBOUR, P_BACKGROUND, P_FALSE_HIT, and P_FALSE_MISS. All possible combinations of the values given in the first two columns of Table 6 were investigated. Results are summarized in the usual format. The standard errors of all the average N values fall in the range 0.2−0.3, this being based on the assumption that the standard deviation of N from a single 20 000-cycle simulation is 1.52 (see above). Table 7 gives the results of the individual simulations with P_FALSE_MISS = P_FALSE_HIT = 0.1.

Several of the results given earlier are reinforced by these high-precision simulations. Thus, N rises as P_FALSE_HIT

Table 6. High Precision (20 000 Cycle) Simulations[a]

parameter	value	N(av) random	N(av) CS	N(av) SE
P_BACKGROUND	0.005	436.6	433.6	454.5
	0.010	437.2	433.5	452.3
	0.025	441.5	438.5	456.4
	0.050	449.7	447.0	461.9
P_NEIGHBOUR	0.05	490.9	490.0	492.7
	0.10	472.8	472.5	479.2
	0.25	429.0	426.6	447.3
	0.50	372.3	363.5	405.8
P_FALSE_MISS	0.0	433.0	428.7	449.8
	0.1	438.7	435.7	454.4
	0.3	452.0	450.1	464.6
P_FALSE_HIT	0.0	422.3	419.7	441.9
	0.1	437.4	435.1	454.4
	0.3	464.0	459.7	472.5

[a] Remaining parameters fixed at SIM_THRESHOLD = 0.80, P_FUNCTION = STEP, N_BATCH = 10, PRE_DELAY = 1, POST_DELAY = 1, N_POSFEED = 10, POS_STRAT = SINGLE, and P_NEGFEED = 0.995.

Table 7. Results of Individual Simulations at Constant Screen Precision[a]

P_NEIGHBOUR	P_BACKGROUND 0.005	0.010	0.025	0.050	
0.05	483	484	491	499	random
	482	484	488	500	CS
	488	489	493	498	SE
0.10	461	464	469	482	random
	465	467	469	483	CS
	476	472	476	483	SE
0.25	413	420	420	430	random
	413	416	418	429	CS
	439	442	445	448	SE
0.50	361	353	354	366	random
	348	349	350	359	CS
	401	395	398	400	SE

[a] Body of table gives N values rounded to nearest integer (standard deviation of individual value = 1.5). P_FALSE_MISS = P_FALSE_HIT = 0.1.

and P_FALSE_MISS are increased, particularly the former; CS affords, at most, only small improvements over random initial ordering, and SE is still worse than random. As expected from earlier simulations, N is highly dependent on P_NEIGHBOUR and P_RATIO. In addition, it can now be seen that N varies only slowly with P_BACKGROUND, though such variation as exists is in the expected direction, i.e., P_BACKGROUND and N increases together.

Analysis of Compound Selection. The final series of simulations focussed on the number of compounds that are promoted in the order due to positive feedback or relegated by negative feedback. Sixteen sets of model parameters were chosen (Table 8) and three simulations performed for each, using set 1 and 2500 cycles, and with the initial order of testing chosen by CS, SE, or at random. In each simulation, a count was made of the number of compounds that were promoted or relegated through the order. These figures are given in Table 8, expressed as a percentage of the total number of compounds selected for screening.

Typically, between 10−25% of tests are performed on compounds selected by positive feedback. As would be expected, the percentage is toward the top end of this range when P_RATIO, P_BACKGROUND, and N_POSFEED are high. There is a small but probably genuine tendency for more compounds to be promoted through positive feedback when the initial order of testing is chosen by CS rather than at random. This is particularly true at high values of P_RATIO and P_BACKGROUND. One possible reason is that CS tends to concentrate on large clusters first, where more near neighbors are available should a weakly active molecule be found. Conspicuously fewer compounds are chosen by positive feedback when the initial order is based on the SE algorithm. This is discussed further in the following section.

Negative feedback has a negligible effect on compound selection when P_NEGFEED = 0.8. However, a large number (20−45%) of compounds are relegated when P_NEGFEED is close to unity. This is particularly true when the initial order is selected at random. The probable reason is that the CS algorithm starts by selecting just one compound from each cluster. During this phase, negative feedback cannot be operational, since none of the compounds has any tested near neighbors. SE tends to select outliers, which have no near neighbors, tested or otherwise.

CONCLUSIONS

In most of the simulations described above, the highly active molecule was found in less than 500.5 tests. To this extent, the computer algorithms "worked", i.e., produced an improvement in hit rate compared with what might have been expected at random. However, this simple conclusion must be qualified in a number of ways. Firstly, the improvements were often small, particularly when P_RATIO was low. This means that the computer algorithms were ineffective when there was only a weak structure−activity relationship in the vicinity of the highly active molecule. Secondly, almost no gains were achieved for the more diverse set of test molecules, set 2. Thus, the algorithms struggle when applied to sets of compounds that contain few clusters. Thirdly, such improvements as were obtained were due almost entirely to the feedback mechanisms. Arranging the compounds into an initial order based on cluster sampling was only mildly effective compared with random initial ordering. Selecting the initial order by stepwise elimination (i.e., maximum dissimilarity) was actually counterproductive.

The disappointing performance of the cluster sampling and maximum dissimilarity algorithms is the most important result of this study, particularly given the current interest in such techniques.[8] Insight can be gained by considering a hypothetical example. Suppose that five compounds are available, of which two (A and B) form a cluster, two (C and D) form a second cluster, and one (E) is an outlier. Suppose, further, that one of the compounds is very active; that, if it is a member of a cluster, its near neighbor is weakly active (i.e., a structure−activity relationship exists); and that all other compounds are inactive (i.e., the background incidence of weak activity is zero). In choosing the initial order of screening, the experimentalist has two basic choices: to screen the outlier first, or a member of a cluster. The former strategem would be exemplified by the order E, A, C, B, D and the latter by A, C, E, B, D. Table 9 summarizes all possible outcomes for each of these initial orders, assuming that positive feedback will be used when and as soon as appropriate. It shows that the highly active molecule is likely to be found more quickly if the outlier is *not* screened first.

Table 8. Percentages of Compounds Promoted or Relegated through Feedback[a]

simulation parameters				random		CS		SE	
P_BACKGROUND	P_RATIO	N_POSFEED	P_NEGFEED	prom[b]	releg[c]	prom[b]	releg[c]	prom[b]	releg[c]
0.05	1.0	5	0.800	17.0	3.5	17.0	1.1	9.6	1.1
0.05	1.0	10	0.800	21.9	3.5	21.8	1.1	11.4	1.1
0.01	5.0	5	0.800	14.3	0.0	13.9	0.0	7.8	0.0
0.01	5.0	10	0.800	19.0	0.0	18.2	0.0	9.3	0.0
0.05	10.0	5	0.800	18.3	2.8	20.8	0.6	9.4	0.4
0.05	10.0	10	0.800	24.2	2.6	27.0	0.4	10.8	0.3
0.01	50.0	5	0.800	15.4	0.0	16.8	0.0	7.4	0.0
0.01	50.0	10	0.800	20.6	0.0	23.1	0.0	8.8	0.0
0.05	1.0	5	0.995	12.0	45.9	15.1	30.7	7.9	26.6
0.05	1.0	10	0.995	15.0	45.5	19.0	31.1	9.4	27.4
0.01	5.0	5	0.995	11.4	43.3	13.5	28.6	7.3	23.3
0.01	5.0	10	0.995	14.5	42.1	17.4	28.9	8.9	24.5
0.05	10.0	5	0.995	14.4	40.4	18.6	23.9	8.3	19.4
0.05	10.0	10	0.995	17.9	38.2	24.6	21.7	9.9	18.2
0.01	50.0	5	0.995	12.6	37.6	16.7	21.0	7.2	16.3
0.01	50.0	10	0.995	16.2	36.4	22.2	20.2	8.5	15.0

[a] Other simulation parameters fixed at SIM_THRESHOLD = 0.80, P_FUNCTION = STEP, P_FALSE_MISS = 0.1, P_FALSE_HIT = 0.1, N_BATCH = 10, PRE_DELAY = 1, POST_DELAY = 1, and POS_STRAT = SINGLE. [b] Prom = percentage of compounds promoted through positive feedback. [c] Releg = percentage of compounds relegated through negative feedback.

The key event in Table 9 is the detection of weak activity in the close structural analogue of the very active molecule. It is important because it prompts a change in the order in which the remaining compounds are tested which, in turn, causes the highly active molecule to be found more quickly. In the example, this is the *only* mechanism available for improving the odds and *can only be applied within a cluster.* Testing the outlier first is therefore undesirable because it reduces the probability of such positive feedback occurring early in the screening process. Therein lies the reason for the poor performance of SE: it lowers the hit rate by placing a high proportion of outliers at the top of the order, thereby delaying the beneficial use of positive (and negative) feedback. This, of course, is consistent with the data presented in Table 8.

Given this analysis, it is puzzling that CS is also relatively ineffective (though at least not counterproductive). One possible reason is that the algorithm is suboptimum. For example, it starts by selecting compounds from large clusters, whereas it may be preferable to sample the smaller clusters first. Also, the algorithm starts by selecting only one compound from each cluster, whereas it might be preferable to choose more. Thus, the possibility exists that a better cluster sampling algorithm could be devised, which would given substantial improvements in hit rate. In contrast, a more effective maximum dissimilarity algorithm than SE (i.e., one that more accurately locates and selects outliers) is almost certain to be even more counterproductive. The hypothetical example also suggests why it was difficult to improve the hit rate for set 2. This set of compounds contains few clusters and many outliers, so the gains achieved by feedback were bound to be limited. It is interesting that CS was more useful for set 2 than for set 1, implying that it is more important to seek out clusters by algorithmic methods when they are rare.

The crucial importance can now be seen of the assumption made at the beginning of this paper: that, in a set of random compounds, each has an equal a priori probability of being active. If it is true, maximum dissimilarity methods will tend to lower hit rates by reducing the effectiveness of feedback. But if it is false—specifically, if outliers are inherently more likely to be active—then maximum dissimilarity may be desirable rather than counterproductive.[1] Arguments to support an elevated probability of activity amongst outliers can be formulated easily. For example, natural products are especially likely to show biological activity and are also likely to be outliers. On the other hand, it is just as easy to

Table 9. Possible Outcomes of Hypothetical Screening Experiment[a]

(a) Initial Order E, A, C, B, D

highly active molecule	test-by-test outcome of screening: test no.	compd tested	result of test	action	NTEST
A	1	E	inactive	none	
	2	A	highly active	finish	2
B	1	E	inactive	none	
	2	A	weakly active	promote B	
	*3	B	highly active	finish	3
C	1	E	inactive	none	
	2	A	inactive	none	
	3	C	highly active	finish	3
D	1	E	inactive	none	
	2	A	inactive	none	
	3	C	weakly active	promote D	
	*4	D	highly active	finish	4
E	1	E	highly active	finish	1
				av NTEST = 2.6	

(b) Initial Order A, C, E, B, D

highly active molecule	test-by-test outcome of screening: test no.	compd tested	result of test	action	NTEST
A	1	A	highly active	finish	1
B	1	A	weakly active	promote B	
	*2	B	highly active	finish	2
C	1	A	inactive	none	
	2	C	highly active	finish	2
D	1	A	inactive	none	
	2	C	weakly active	promote D	
	*3	D	highly active	finish	3
E	1	A	inactive	none	
	2	C	inactive	none	
	3	E	highly active	finish	3
				av NTEST = 2.2	

[a] Summary of all possible outcomes of the hypothetical screening experiment described in the text (i.e., the outcomes assuming each compound in turn is highly active). Asterisks indicate tests performed on compounds promoted in the order through positive feedback. NTEST = number of tests performed to find highly active molecule.

argue the opposite contention. For example, clusters in a typical company's compound collection may often correspond to synthetic series that have been made with due respect to chemical and metabolic stability. Overall, it is difficult to reach any sort of reliable conclusion. Another, more important, complication is that a premium may be placed on detecting activity in outliers (see Introduction); indeed, this is rather likely. In this event, the experimentalist may choose to use maximum dissimilarity methods to optimize the chances of finding "novel" hits but at the price of lowering the *overall* hit rate.

Finally, some comments are called for on the practical importance of this work given the advent of combinatorial chemical libraries and very high-throughput biochemical assays.[9] The development of these techniques makes it possible to screen millions of compounds, which, by the very nature of a combinatorial library, will be uniformly distributed in a particular molecular-structure space (i.e., there will be no outliers). At first sight, it might therefore appear that the traditional random screening of in-house, historical collections of synthetic compounds will be entirely superseded. However, this is unlikely to be the case. Firstly, it is especially valuable to find activity in an in-house compound, which may be proprietary and designed to be chemically and metabolically stable. Secondly, very high throughput rates cannot be achieved with all biological assays. Thus, the traditional random screening of in-house collections of synthetic compounds is likely to remain an important method of lead generation.

A forthcoming paper (in preparation) will describe how a modified cluster sampling and feedback approach has substantially increased hit rate in one of our in-house assays.

ACKNOWLEDGMENT

Peter Willett (University of Sheffield, UK) is thanked for helpful comments. Zeneca Agrochemicals in the UK is part of Zeneca Limited.

REFERENCES AND NOTES

(1) Lajiness, M. S. An evaluation of the performance of dissimilarity selection. In *QSAR: Rational Approaches to the Design of Bioactive Compounds*; Silipo, C., Vittoria, A., Eds.; Elsevier: Amsterdam, 1991; pp 201−204.
(2) Willett, P.; Winterman, V.; Bawden, D. Implementation of non-hierarchical cluster analysis methods in chemical information systems. Selection of compounds for biological testing and clustering of substructure search output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109−118.
(3) *Concepts of Molecular Similarity Analysis*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990.
(4) Molecular Access System; Molecular Design Limited, California, U.S.A.
(5) Willett, P.; Winterman, V. A comparison of some measures for the determination of inter-molecular structural similarity. *Quant. Struct. Activ. Relat.* **1986**, *5*, 18−25.
(6) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared nearest neighbors. *IEEE Trans. Comput.* **1973**, *C-22*, 1025−1034.
(7) Spath, H. *Cluster Analysis Algorithms*; Ellis Horwood: Chichester, UK, 1980.
(8) *Proceedings of the First Forum on Data Management Technologies in Biological Screening*; Carter, C., Freter, K. R., Eds.; SRI International: California, 1992.
(9) Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries. *J. Med. Chem.* **1994**, *37*, 1233−1251.

CI9400775