Table VII. P388 Results on Certain of the 988 Compounds Which Were Candidates for Elimination by Computer Selection

	308 low active[a]	305 high occur[b]	agree[c]	combined 488[d]	all 988
confirmed active	3	2	1	4	26
presumptives	7	7	2	12	33
pass/fails	4	4	2	6	10
toxic	5	8	1	12	27
$A + 0.6P + 0.2(P/F) + 0.1T$	8.5	7.8	2.7	13.6	50.5

[a] These 30% compounds scored in the lowest 10% according to the range of the P388 training set. [b] These 305 compounds met the criterion for adequate representation in P388 testing. [c] Agreement or overlap between the compounds in the first two columns. [d] All the compounds that satisfied either or both of the first two columns.

Table VIII. Agreement between the 308 Low Activity Prediction and the 305 High Occurrence Compounds As Defined in Table VII

	occurrence		
activity	high	other	total
low	125	183	308
other	180	500	680
total	305	683	988

just 125 compounds; see Table VIII. This agreement quantifies to $\kappa = 0.142$, about the same as the 0.145 agreement between the computer and the chemist in their choices. Such an unforeseen, albeit low agreement, corresponds to the unexpectedly small number of actives among the 305 high occurrence compounds.

Thus, it is expected that compounds will be eliminated if they yield low activity score and/or high occurrence in the training sets. In the current study, this would exclude 488 compounds or almost half of the 988 compounds. The 488 compounds contain only 4 of the 26 confirmed actives, 12 presumptives, 6 pass/fails, and 12 toxic compounds for an estimated 13.6 expected confirmed actives. The remaining 500 compounds would have 36.9 expected confirmed actives for an enrichment of 36.9/25.6 or 1.44 with a standard deviation of 0.109. So there are 4.0 standard deviations for statistical significance with a P value below 0.0001. This compares with the enrichment of 1.34 and P value of 0.0022 achieved by selection of the top half by activity alone.

Again, the enrichment should be even better since the presumptives, etc., of the eliminated compounds have been given the same weight as their counterparts among the selected compounds. But the experience of actual confirmed actives indicates that the ultimate confirmed actives will continue to bias in favor of the selected compounds.

Early experience on the use of an operational model installed April 1979 to rate potential acquisitions continues to reinforce the trends found in the study just reported. When more biological data is accumulated this operational experience will be reported. Beginning March, 1980, the computer began an active participation in selection of compounds for screening in P388 from all potential acquisitions.

ACKNOWLEDGMENT

Study II owes its inception to Sidney Richman. The essential chemistry was performed by Kenneth Paull who was always encouraging. The successful conclusion was wrought by statistical help from Robert Tarone. Alfred Feldman and Ruth Geran read the manuscript and provided helpful criticism and advice.

REFERENCES AND NOTES

(1) Hodes, L.; Hazard, G. F.; Geran, R. I.; Richman, S. "A Statistical-Heuristic Method for Automated Selection of Drugs for Screening". *J. Med. Chem.* **1977**, *20*, 469-475.
(2) Hodes, L. "Computer-Aided Selection of Novel Antitumor Drugs for Animal Screening". *ACS Symp. Ser.* **1979**, *No. 112*, 583-602.
(3) NSC numbers are assigned in chronological order to compounds sent for biological testing. NSC 260 000 corresponds to the end of 1975.
(4) Similar difficulties with singly occurring features were reported in a previous work (Cramer, R. D., et al. *J. Med. Chem.* **1974**, *17*, 533-535) which was a precursor of the current work.
(5) Reference 2, p 598-600.
(6) Many compounds (2 or 3%) of the NCI file are sent to biological testing with partially or wholly undefined structure.
(7) The AA keys are described in ref 1 and the gAA keys are described in ref 2, p 597-598. All the computer results in Tables I-VIII are derived from the gAA version.
(8) Fleiss, J. L., "Statistical Methods for Rates and Proportions", Wiley: New York, 1973.
(9) The T/C ratio, in P388, is the median life span of the treated animals divided by that of the controls.
(10) Lehmann, E. L., "Nonparametrics: Statistical Methods Based on Ranks"; Holden Day: San Francisco, 1975.

Selection of Molecular Fragment Features for Structure-Activity Studies in Antitumor Screening

LOUIS HODES
National Cancer Institute, Silver Spring, Maryland 20910

Received May 4, 1981

The National Cancer Institute Developmental Therapeutics Program screens about 13 000 compounds per year for antitumor activity in a mouse prescreen (P388). A method for predicting activity in P388 uses molecular fragment features of potential acquisitions. This paper covers some details about how the set of features was chosen, filling a gap in earlier publications.

INTRODUCTION

Molecular structure fragments offer a rich body of features for the prediction of biological activity. One such predictive method[1-3] is being used to aid in the selection of compounds for antitumor screening in the National Cancer Institute Developmental Therapeutics Program (NCIDTP). An in vivo prescreen, mouse lymphocytic leukemia or P388, is used for antitumor testing of large numbers of compounds, about 13 000 per year.

Two or three times that number of compounds are collected in the form of structure diagrams as potential acquisitions. These are searched for duplicates in our chemical information system.[4] At the same time the structures are run through the following programs which help in the acquisitions process.

These programs use structure features to estimate the probability of activity in P388. Further, any such features not yet seen in P388 testing are flagged as unique. Also, the incidence of the least occurring feature is used as an indicator of whether the compound is already well represented in P388 testing.

By examining the structures in conjunction with the activity and novelty estimates generated by these programs, a medicinal chemist reduces the potential acquisitions to the 13 000 acquisitions.

The structure features used in this preselection subsystem are molecular fragments that were developed for substructure search of the NCIDTP file. Although several other collections of fragments were tried, the availability and suitability of the NCIDTP fragments led to their use on potential acquisitions.

Earlier reports describe the method for estimating a measure of the probability of activity,[1] the use of the method on the large volume of NCIDTP data,[2] and, most recently, a report on validation of the method.[3] Missing from the earlier work[2] was a statistical analysis of the difference in performance of two sets of features. This analysis, presented here, provides some insight into the nature of the chemical structure data. This work should be applicable to other areas where descriptors are used to represent data.

NCIDTP DATA

The method for predicting activity uses the incidence of each feature in compounds of known activity to compute an activity weight for the feature. The collection of compounds of known activity is called a training set. Upon completion of the training phase, i.e., the computation of activity weights, these weights are applied to compounds of presumably unknown activity. The unknown compounds can be referred to collectively as a test set.

The data for this exercise consisted of a training set obtained from a search of the biological file early in 1977. The search was restricted to NSC numbers below 260 000. NSC numbers are assigned upon acceptance of a compound for biological testing. NSC number 1 was assigned in 1955 and number 260 000 corresponds to the end of 1975, so biological testing should have been fairly complete at the time of the search. There was not yet an enormous amount of data since P388 did not become the standard prescreen until early in 1976.

Compounds were collected at three levels of activity based on T/C, the ratio of the median life span of the treated animals to that of the controls. The highly active compounds (A's) had T/C of at least 175% in two separate experiments. The moderately active compounds (C's) had T/C of at least 120% in two separate experiments but were not in category A. The inactive compounds (N's) had T/C always less than 120% including an experiment with at least three dose levels, the lowest two nontoxic.

The biological search was followed by a series of substructure searches to remove classes of well-known compounds. This follows Cramer et al.[5] in using the method to find new leads. Also, the large number of highly active analogues would exaggerate performance while, at the same time, overweight fragments that occur together with highly active moieties.

The training set consisted of all compounds in the three categories just mentioned minus the compounds from familiar classes; also removed were those compounds which lacked a well-defined structure. The process of collecting this training set has been described in more detail elsewhere.[2] There were about 85 A's, 1120 C's, and 13 500 N's in the training set.

Two test sets were selected on the basis of NSC ranges. The one that will be pertinent was that chosen from the range of NSC numbers between 268 001 and 272 000 as having relatively few compounds from the well-known classes compared

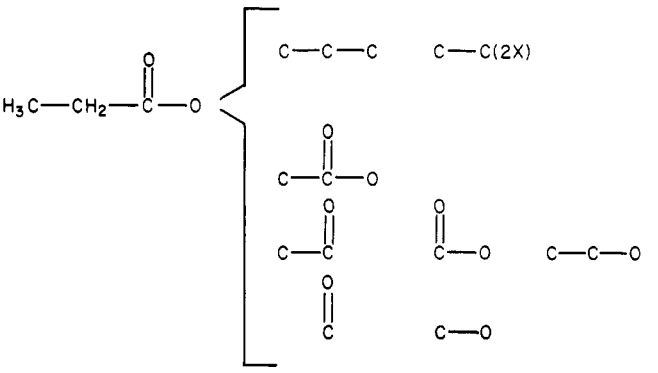

Figure 1. A compound and its AA fragments; each nonterminal nonhydrogen atom is taken as a central atom.

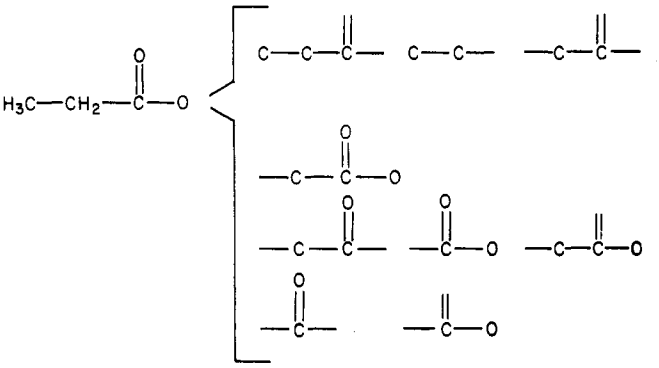

Figure 2. gAA fragments for the compound of Figure 1; each AA fragment yields a gAA fragment.

to adjacent ranges of NSC numbers. The biological data was obtained according to the activity categories mentioned earlier by means of a manual review of the screening data summaries. There were 32 A's and 145 C's in the test set.

Almost all the A's (28) and 77 of the C's belonged to the well-known classes removed from the training set. The 68 remaining C's were considered more important to our program as relatively novel. These compounds were relabeled G.

Since the test set came from a limited range of NSC numbers, there was some clustering of similar structures, which limited diversity. Therefore, results on these active compounds were not the only factor in the choice of features, as discussed elsewhere.[2]

NCIDTP STRUCTURE FEATURES

Two kinds of structure keys are used for substructure search in the NCIDTP chemical information system. The first kind are the ring and nucleus keys. These are generally useful but will not be very pertinent to the discussion here. The second kind are small atom-centered fragments called augmented atom (AA) keys. Like the ring keys the AA keys are generated exhaustively for each structure, as shown in Figure 1.

The other relevant set of features are extensions of the AA keys called the ganglia AA or gAA keys. These fragments are illustrated in Figure 2. Each one consists of an AA fragment plus the additional bonds on all the atoms. Thus, for each structure the total number of AA and gAA keys are the same, but there may be several more distinct gAA keys than there are distinct AA keys. This depends on the multiplicity of the AA keys since two identical AA keys may extend to different gAA keys. There is an example of such a key in Figures 1 and 2.

The gAA keys are intended as a backup in case the AA keys were not discriminating enough for substructure search of the relatively large NCIDTP file. To be used for that purpose the gAA fragment set for each compound would need to be

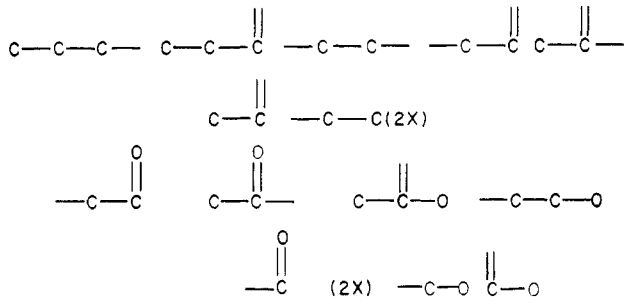

Figure 3. Additional fragments for the same compound as they would be required for substructure search by gAA keys.

Table I. Ranking of Active Compounds in Deciles by the AA Model

rank, deciles	no. of A compd	no. of G compd	no. of C compd
10	18	20	22
9	4	8	9
8	2	6	11
7	2	6	10
6	0	4	8
5	0	6	3
4	1	7	8
3	3	2	2
2	2	7	3
1	0	2	1

Table II. Ranking of Active Compounds in Deciles by the gAA Model

rank, deciles	no. of A compd	no. of G compd	no. of C compd
10	17	22	35
9	1	13	12
8	2	3	7
7	3	3	5
6	2	6	3
5	0	6	3
4	4	5	3
3	3	2	2
2	0	5	0
1	0	1	3

expanded to include additional fragments with subsets of ganglia bonds as shown in Figure 3. This expansion is, of course, unnecessary when the fragments are used for structure–activity studies. Although the gAA keys have never been used for substructure search, their existence was serendipitous for the structure–activity determinations.

It is clear that any large diverse set of compounds will contain several times more gAA keys than AA keys since several bond variations are possible on each AA key. The active compound of the training set, numbering about 1200, had about 600 ring and nucleus keys and 1200 AA keys. They also had about 7100 gAA keys; so in going from the AA to the gAA model the number of features increased from 1800 to 7700.

The performance of the AA and gAA models on the test data described earlier is shown in Tables I and II, respectively, repeated with minor corrections from ref 2. In the next section the performance will be analyzed in some detail. It is noted here that the AA and gAA fragment sets seemed to represent two levels of performance as standards for other feature sets. This remark will be amplified in the succeeding section.

AA VS. GAA

Tables I and II indicate that on the highly active A's the AA model performed slightly better than the gAA. However, the gAA model was overwhelmingly better on the C's while showing only slightly better on the G's.

Table III. Comparison of Ranking of the 32 A Compounds between the AA and gAA Models

NSC no.	AA rank	gAA rank	diff	diff midrank
268223	2459	2131	328	25
268238	3233	3218	15	12.5
26840	3202	3209	−7	7
268241	3206	3197	9	8.5
268242	3219	3215	4	3.5
268704	3231	3194	37	18
268708	3195	3205	−10	10
268710	3234	3219	15	12.5
268858	2139	2360	−221	22
268927	566	803	−237	24
268958	2707	2239	468	29
268965	336	675	−339	26
268966	2464	2255	209	20
268970	847	1075	−228	23
268985	1095	1127	−32	17
269043	853	837	16	14
269086	883	1282	−399	28
269148	3225	3248	−23	15
269150	3237	3251	−14	11
269433	3207	3201	6	6
269434	3203	3204	−1	1
269608	2826	1914	912	32
269609	2035	2541	394	27
269610	3144	2930	214	21
269728	2615	1895	720	31
269754	3260	3256	4	3.5
269757	3253	3257	−4	3.5
269760	3254	3258	−4	3.5
270515	3018	3128	−110	19
270516	3042	3051	−9	8.5
271882	1989	1306	683	30
271936	3232	3208	24	16

First these results will be quantified so that the significance of the difference can be evaluated. Then some reasons will be given for the pecularities in the results. Recall that the results on the test set under consideration were not the main reason for selecting the gAA model in preference to the AA model. As shown in ref 2 and verified in ref 3, results on larger data sets were more conclusive.

The AA and gAA models will be compared by their performance in ranking the active compounds according to the respective activity scores. That is if one model consistently ranks the same active compounds higher than the other model ranks them, the former model is thereby shown to be superior. Note that the activity scores themselves are not necessary for this comparison; we need only the rank ordering.

For an illustration, Table III lists the highly active A's by NSC number together with their rankings by the AA and gAA models. These rankings are directly comparable since the same test set was used in both cases. The differences are listed in the third column of Table III. They are positive if the AA rank is larger and negative if the gAA rank is larger. Tied ranks would yield zero difference but there were not any in these data.

The differences can be subjected to several tests. The simplest is a sign test. That is, the chance of occurrence of the positive and negative differences are tested as though they were heads and tails in a coin toss. A more sophisticated test takes into account the magnitude of the differences by ranking them regardless of sign and applying the Wilcoxon signed rank test. This test and the sign test are thoroughly described in Lehmann.[6]

For example, there are 17 positive and 15 negative differences among the A's in Table III. This is almost an equal split. Using the sign test, involving a normal approximation with

Table IV. Comparison of AA and gAA Ranking on the Active Compounds for the Sign Test

	A	C	G	total
AA > gAA (+)	17	23	35	75
AA < gAA (–)	15	54	33	102
no. of SD	0.18	3.4	0.12	2.03
P value		0.0007		0.04

Table V. Comparison of AA and gAA Ranking on the Active Compounds for the Wilcoxon Signed Rank Sum Test

	A	C	G	total
rank sum (+)	309.5	798.5	1024	6095
rank sum (–)	218.5	2204.5	1322	9658
no. of SD	0.85	3.6	0.91	2.6
P value		0.0004		0.009

continuity correction, the result is only 0.18 of a standard deviation (SD) away from the expected value.

The magnitudes of the differences are themselves ranked in the last column of Table III. In cases of ties in magnitude, the average or midrank is used. Note that the four highest ranks belong to positive differences. This should give more significance to the Wilcoxon test. The rank sums are 309.5 for the positive differences and 218.5 for the negative differences. The expected value is 264, half the total of 528, and the SD works out to be 53.5. So the normalized unbalance of the differences is (309.5 – 264)/53.5 or 0.85. This is still far from statistical significance, but it is quite different from the simple sign test.

If we consider all the actives together, the situation is reversed and seems quite decisive. On the entire set of 177 active compounds, the gAA model ranked 102 higher compared to 75 for the AA model. The sign test alone yields a statistical significance at 2.03 SD, *P* value of 0.04. The Wilcoxon test increases this lead to 2.6 SD and a *P* value of 0.009.

Unfortunately, neither the highly active A's described earlier nor the relatively novel G's show results similar to the entire set of actives. Indeed, in the G's as well as the A's the sign test slightly favors the AA model. There are more positive differences than there are negative among the G's despite the opposite showing in Tables I and II. Of course, the margin of 35 to 33 is even less significant than the small fraction of a SD in the case of the A's. Also, the more precise Wilcoxon test does show some advantage for the G's under the gAA model, although this is not significant at 0.91 SD.

The reason for the big overall gAA lead lies mainly with the 77 C's that were analogues of well-known active compounds. These showed exceptionally well under the gAA model despite the removal of similar compounds from the training set. There were 54 compounds of the 77 that ranked higher on the gAA than they did on the AA model. By the sign test alone, this shows 3.4 SD with a *P* value of 0.0007. The Wilcoxon test increases the margin slightly to 3.6 SD or a *P* value of 0.0004. The results are summarized in Tables IV and V.

DISCUSSION

It should be apparent to the reader that there are some anomalies from a statistical point of view. If the gAA model is so much better than the AA model as indicated by its performance on all the active compounds, then the set of 32 A's and, most certainly, the set of 68 G's should have been sufficient to show preference for the gAA model.

One can dismiss these discrepancies by merely pointing out that selection of the compounds is not statistically random. Each set tends to cluster groups of similar compounds; this has the effect of a reduction in sample size. However, there are other, more pertinent, reasons for the discrepancies.

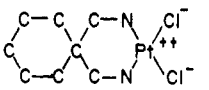

Figure 4. This highly active compound scored highest on the AA model with respect to the gAA model. Although the platinum compounds were removed from the training set as a well-known class, the AA key consisting of a metal connected to two halogens has an activity weight of 1.25. The corresponding gAA key for this compound does not occur in the training set.

Such a reason is uncovered by a look at NSC 269608 in Figure 4. This was the A compound that did best in the AA ranking relative to the gAA ranking. There is an AA key with high activity weight such that the corresponding gAA key in this compound does not occur in the active training set. Thus, although the gAA key is more discriminating, it can be too specialized in some cases and can be outperformed by the AA key.

This failure becomes less of a problem as the training set increases in size so that there are fewer gAA keys that have not yet appeared. Another example of this phenomenon occurs in the work of Tinker,[7] who applied the same method with equivalent molecular fragment features to estimate mutagenicity. Tinker notes[8] that the AA fragments had a slight edge when his training set was 700 compounds, but the gAA fragments proved superior with an increase to 800 compounds.

There are two other, more complete, solutions to this problem. One which is used at NCIDTP is simply to flag as unique all gAA keys which have not appeared in the training set. These are independently of special interest since the screening program is intended to cover the widest possible range of compounds.

Another method of correcting for the discrepancy would be to combine the AA and gAA models. This was not done because of the redundancy between the sets of features—the method ideally requires independent features. Such redundancy can be eliminated by using conditional probabilities. That is, for each gAA key one should use incidence relative to the incidence of the underlying AA key. Such a scheme may be tried in the near future. It is complicated by the use of multiplicities of features.

FURTHER DEVELOPMENTS

Figures 1–3 depict molecular fragments as they are generated for substructure search. All subfragments that have the form of legitimate fragments are required so that a query containing the subfragment would appropriately retrieve the structure. However, there would be enormous amount of redundancy if such fragments were to be used for structure–activity determinations.

At the time of ref 1, redundancy in the set of AA keys was reduced by the expedient of eliminating AA fragments with an even number of atoms; i.e., essentially only triples and five atom keys were kept. This reduction in redundancy yielded a noticeable boost in performance. With the gAA keys a similar approach was taken at first though in this case the performance was not affected. The previous section comparing the performance of the two sets of keys included such reductions in keys. However, it was soon uncovered that there was a more natural way to eliminate redundancy within the gAA keys.

It was possible to obtain a single, largest gAA key centered at each atom by merely ignoring those gAA keys which have ganglia (extra bonds) on the central atom.[9] Thereby, the molecule of Figure 1 would have only the fragments shown in Figure 5. Later tests showed the reduced set of gAA keys to perform better, especially at the upper end of the ranking.

Since there is no indication as to whether any given AA key is the largest at its central atom, there is no equally direct way

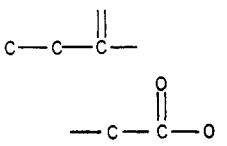

Figure 5. Largest gAA key at each nonterminal nonhydrogen atom for the compound of Figure 1.

to obtain the largest AA key. The easiest way to achieve this from the NCIDTP system is to generate the largest gAA keys as just described, drop the ganglia to get the underlying AA keys, and merge to count multiples of identical AA keys. One would certainly use such a method when combining AA and gAA models.

Several other sets of fragments were tried during the course of this work. The first was a set developed at the Walter Reed Army Institute of Research (WRAIR) for substructure search.[10] The methods[11] employed to generate the WRAIR fragments allowed superior elimination of redundancy.

On the small data set of ref 1, these fragments gave poor results. The difficulty was lack of bond specificity. Although the fragments included sizes of linear sequences to 7 atoms and ring-chain hybrid fragments to 11 atoms, there was no bond specificity, even at the size of the AA keys.

Another interesting set of fragments were made available by the Basel information Center for Chemistry (BASIC).[12] These fragments are routinely generated for all CAS registrations; so they were intended to be used in a model for literature surveillance. The BASIC keys include linear sequences of four to six atoms, with bonds specified only as to their ring or nonring character. These keys performed at about the same level as the AA keys.

The gAA keys may be prohibitively expensive to generate for all CAS registrations. Therefore, a new set of features was designed so that they could be quickly generated from the CAS standard distribution format. These are similar to the gAA keys, but are bond centered instead of atom centered.

The bond-centered (BC) fragments were first developed as an approximation of functional groups.[13] In the following a bond will be defined to include the atoms at both ends. Let us call a bond admissible if it is not a carbon–carbon single bond or a carbon–carbon ring alternating bond. An algorithmic definition of a functional group will be maximally connected set of admissable bonds. These functional groups were tried as fragment features and were found to be too specific. In the NCIDTP data described earlier, about 20% of the compounds in the test set did not have any features from the training set.

To reduce their specificity the BC fragments were defined as follows. Each admissible bond forms the nucleus of a BC fragment. The BC fragment consists of the admissible bond and all its adjoining bonds. Also, when a nonadmissible bond is flanked by admissable bonds at both ends, these all together form a BC fragment. The BC fragments, generated at the rate of about 50 compounds/s (370/168 CPU time), performed about as well as the gAA keys which take 1 s for about three compounds.

One advantage of the gAA keys over the AA keys lies in the treatment of terminal atoms, atoms which have only one nonhydrogen neighbor. In the gAA fragment a terminal atom is distinguished by having no ganglia, but there is no such distinction in the AA fragment. The additional information about terminal atoms accounts for some of the superior performance of the gAA keys. This property of gAA fragments carries over to some extent to BC fragments since a terminal admissible bond is distinguished.

CONCLUSION

Predicting activity over a broad range of compounds is fraught with difficulty. The Hansch-type methods using physicochemical parameters do not seem to work over a variety of classes.[14] Still, activity is determined by structure. Ideally, features should be three-dimensional configurations. However, these would be too varied and expensive to compute for large numbers of compounds, especially in this application where mechanisms and receptor sites are, for the most part, undetermined.

A recent study[3] has shown the method used here to compare favorably with selection by a chemist familiar with the data. This paper, by examining one aspect of the method is some detail, brings out some limitations. These are due to problems with the biological data, chemical structures, choice of fragments as features, the method itself, and difficulty of the application. It is this last consideration that has led to the usefulness of this work.

REFERENCES AND NOTES

(1) Hodes, L.; Hazard, G. F.; Geran, R. I.; Richman, S. "A Statistical-Heuristic Method for Automated Selection of Drugs for Screening". *J. Med. Chem.* **1977**, *20*, 469–475.
(2) Hodes, L. "Computer-Aided Selection of Novel Antitumor Drugs for Animal Screening". *ACS Symp. Ser.* **1979**, *No. 112*, 583–602.
(3) Hodes, L. "Computer-Aided Selection of Compounds for Antitumor Screening. Validation of a Method". *J. Chem. Inf. Comput. Sci.*, preceding paper in this issue.
(4) Richman, S., et al. "The Drug Research and Development Chemical Information System of NCI's Developmental Therapeutics Program". *ACS Symp. Ser.* **1978**, *No. 84*, 200–221.
(5) Cramer, R. D., III, et al. "Substructural Analysis. A Novel Approach to the Problem of Drug Design". *J. Med. Chem.* **1974**, *17*, 533–535.
(6) Lehmann, E. L. "Nonparametrics: Statistical Methods Based on Ranks"; Holden Day: San Francisco, 1975.
(7) Tinker, J. "Relating Mutagenicity to Chemical Structure". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 3–7.
(8) Reference 7, p 6.
(9) Lefkovitz, D.; Hill, H., private communication.
(10) Feldman, A. P.; Hodes, L. "An Efficient Design for Chemical Structure Searching. 1. The Screens". *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 147–152.
(11) Hodes, L. "Selection of Descriptors According to Discrimination and Redundancy. Application to Chemical Structure Searching". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 88–93.
(12) Schenk, H. R.; Wegmuller, F. "Substructure Search by Means of the Chemical Abstracts Service Chemical Registry II System". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 153–161.
(13) The functional group fragment algorithm, together with a program for generating the fragments, was given to the author by Paul Blower of Chemical Abstracts Service. This program was adapted to generate the BC fragments.
(14) Martin, Y. C. "A Practitioner's Perspective of the Role of Quantitative Structure–Activity Analysis in Medicinal Chemistry". *J. Med. Chem.* **1981**, *24*, 229–237. See p 235.