biological activities, and various types of parameter dependencies. For space reasons other examples where the method could be usefully applied have not been included. In addition to these, some other series were found in the literature where the activity spread between the members of the initial compound group was too small to permit a meaningful analysis.

A key feature is that, based on the results from only four or five readily available analogues, the correct synthetic direction for increased potency can often be determined. At this stage the parameter dependency can usually be narrowed to a small range of possibilities and further substituents can be chosen which should increase potency no matter what the precise activity–parameter relationship is. This marks an important difference between the strategy outlined in the present approach and that of the standard Hansch method. In the latter method the object is to first determine, utilizing a computer based analysis of the results on eight to twelve compounds, a precise activity–parameter relationship in the form of an equation. This equation is then used to select new analogues which should have improved potency. On the other hand, the manual method does not attempt to precisely identify the activity–parameter relationship but seeks to use a more rapidly obtained approximate determination of this relationship as a stepping stone to the identification of more potent analogues.

In terms of numbers of compounds prepared the position reached after preparation of the second compound group in the manual method is roughly equivalent to that reached after a multiple regression analysis on the first group of compounds made in the standard Hansch analysis.[24] Thus, to the extent that the present manual method can successfully narrow the possible operative parameter dependencies at the end of the first stage, it may represent a more advantageous strategy if the primary goal is to find a readily accessible compound in the maximum potency area in the shortest possible time rather than to determine the exact activity–parameter relationship. Also, computers and statistical procedures are not required thus offering greater simplicity of use for most medicinal chemists.

References and Notes

(1) Presented in part at the 167th National Meeting of the American Chemical Society, Los Angeles, Calif., April 1–5, 1974, and the Fifth International Symposium on Medicinal Chemistry, Paris, July 19–22, 1976.
(2) J. G. Topliss, *J. Med. Chem.*, **15**, 1006 (1972).
(3) C. E. Granito, G. T. Becker, S. Roberts, W. J. Wiswesser, and K. J. Windlinz, *J. Chem. Doc.*, 11 (2), 106 (1971).
(4) P. J. Goodford, *Adv. Pharmacol. Chemother.*, 11, 51 (1973).
(5) C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, **86**, 1616 (1964).
(6) C. Hansch, "Drug Design", Vol. I, E. J. Ariens, Ed., Academic Press, New York, N.Y., 1971, p 271.
(7) P. N. Craig, *J. Med. Chem.*, **14**, 680 (1971).
(8) C. Hansch, *Cancer Chemother. Rep.*, **56**, 433 (1972).
(9) C. Hansch, S. H. Unger, and A. B. Forsythe, *J. Med. Chem.*, **16**, 1217 (1973).
(10) F. Darvas, *J. Med. Chem.*, **17**, 799 (1974).
(11) R. Wootton, R. Cranfield, G. C. Shappey, and P. J. Goodford, *J. Med. Chem.*, 18, 607 (1975).
(12) R. T. Buckler, S. Hayao, O. J. Lorenzetti, L. F. Sancilio, H. E. Hartzler, and W. G. Strycker, *J. Med. Chem.*, **13**, 725 (1970).
(13) R. W. Fuller, J. Mills, and M. M. Marsh, *J. Med. Chem.*, 14, 322 (1971).
(14) N. Kakeya, N. Yata, A. Kamada, and M. Aoki, *Chem. Pharm. Bull.*, **17**, 2558 (1969).
(15) A. Cammarata, R. C. Allen, J. K. Seydel, and E. Wempe, *J. Pharm. Sci.*, **59**, 1496 (1970).
(16) C. Silipo and C. Hansch, *Farmaco, Ed. Sci.*, **30**, 35 (1974).
(17) C. Hansch and E. W. Deutsch, *Biochim. Biophys. Acta*, **112**, 381 (1966).
(18) C. Silipo and C. Hansch, *Mol. Pharmacol.*, **10**, 954 (1974).
(19) B. Blank, N. W. Di Tullio, L. Deviney, J. T. Roberts, and H. L. Saunders, *J. Med. Chem.*, 18, 952 (1975).
(20) C. Hansch, K. H. Kim, and R. H. Sarma, *J. Am. Chem. Soc.*, **95**, 6447 (1973).
(21) M. Yoshimoto, K. N. von Kaulla, and C. Hansch, *J. Med. Chem.*, 18, 950 (1975).
(22) G. van den Berg, T. Bultsma, R. F. Rekker, and W. T. Nauta, *Eur. J. Med. Chem.*, **10**, 242 (1975).
(23) B. J. Broughton, P. Chaplen, P. Knowles, E. Lunt, S. M. Marshall, D. L. Pain, and K. R. H. Wooldridge, *J. Med. Chem.*, 18, 1117 (1975).
(24) The option exists at this point to transpose into the standard Hansch approach, although the compound mix may not be optimal.

A Statistical-Heuristic Method for Automated Selection of Drugs for Screening

Louis Hodes,* George F. Hazard, Ruth I. Geran, and Sidney Richman

Division of Cancer Treatment, National Cancer Institute, Silver Spring, Maryland 20910.

Received June 30, 1976

A statistical-heuristic method for selecting drugs for animal screening is developed with molecular structure features as predictors of biological activity. The method is intended to work on large amounts of data over varied structures. A trial of this method on a small data set allows some comparison with more sophisticated pattern recognition methods. Problems connected with interdependence among structure predictors are critical in this method and schemes to eliminate redundancy are reviewed. Alternate sets of structure predictors are considered. The discussion here outlines directions to be taken in the near future.

A major activity of the Developmental Therapeutics Program (DTP) in the Division of Cancer Treatment (DCT), National Cancer Institute (NCI), is the development of new drugs useful in the treatment of human cancer. As one means of identifying leads to such drugs, DTP, which subsumed the Drug Research and Development Program (DR&DP), operates an antitumor screening program that involves the testing of compounds in a variety of animal tumor models. Because of the limited capacity for screening, currently roughly 15 000 synthetic compounds per year, and the almost limitless possibilities for obtaining compounds, many approaches to selecting acquisitions or assigning priorities of testing are being explored.

Some of these approaches involve the use of biological test data from previous acquisitions and chemical structure data to create a system for predicting the biological activity of a new compound by examining its chemical structure.[1,2] Chemical structural parts are obvious choices for prediction parameters because of their clear pertinence and easy availability. At NCI automated files currently contain chemical structure on more than 280 000 compounds and

2 426 000 antitumor test results.[3,4]

Automated selection of drugs, as described here, differs from more standard approaches[5,6] to quantitative structure–activity determinations in that it is intended to apply across a broad range of compounds rather than a single class. Therein lies its challenge and its opportunity. There may be unusual and unnoticed combinations of certain chemical structure features that impart a specific biological activity. It is the ability to detect such combinations of features that renders the computer capable of predicting new different active drugs. At the very least, we can expect an enrichment of the incidence of biologically active compounds among those selected by this method.

The basic statistical-heuristic method is described here, together with the results of an experimental trial on a relatively small set of data. This report is intended as the first in a series that will deal with this statistical-heuristic approach to structure–activity studies on data derived from the DR&DP files. Future reports will contain results for the P388 lymphocytic leukemia model, which represents the crucial test for this method, and the L1210 lymphoid leukemia model. Preliminary results on the P388 model are very encouraging. If the validity of this method is demonstrated by extensive feasibility testing, it will be used in an operational environment.

The method to be described uses the chemical structure data in a straightforward way to predict biological activity. See Cramer et al.[7] for a similar approach. This method is more statistically sophisticated than Cramer's without essential loss of simplicity. An implicit assumption of both methods is independence, but the features are often interrelated or redundant. Therefore, there are sections on the chemical structure features and ways to deal with this complicated issue of interdependence.

The structure fragments we use as features will not always appear in context as they do in the Free–Wilson method,[8] for example. Although the quantitative approach appears superficially similar to the Free–Wilson method in its use of additivity of constants for the structure features, the interpretation is quite different. It will be clear from the derivations that the weights and scores we derive later are to be interpreted as measures of the probability of activity of a compound if it has the given features, not as contributions to activity. This is a subtle difference, but it will be seen that it is the justification for our use of additivity.

The method was chosen because it is capable of handling a large number of compounds and features, and it accommodates the widely disparate incidences of the various structure features. Also, it is a simple approach in which it is easy to see why a given feature receives its weight.

Basic Scoring Method. Like all prediction methods, the process begins with a set of known actives and a set of known inactives. These are called training sets and they are used separately in a similar way to derive, for each feature present, a weight for activity and a weight for inactivity. The activity and inactivity scores for a compound are then found by summing the respective weights of all its structure features. These two scores can then be used to establish "priorities" among a set of unknowns in a manner to be described later.

The active training set is used to derive the activity score as follows. Essentially, the method assigns a weight to each structural feature based on the statistical significance of its frequency of occurrence in the active training set. In this way, each structural feature is assigned a numerical weight estimating the probability of activity in the specified test system of a compound that has the feature.

An important aspect of the method is its use, as a standard, of the incidence of the feature in the total 280 000 compound file which has been tabulated. If this incidence is p and there are n active compounds then np actives would be expected to have the given feature, *under the* assumption *that the feature has nothing to do with activity*. The weight is determined by how far the actual number of actives with the feature differs from its expected value. This method is analogous to that of Cramer et al.[7] where the weight given to a feature is simply the actual number of actives with the feature minus the expected number.[9] Our method assigns weights according to the statistical significance of this difference.

It is easiest to illustrate the computation of the weight by an example from a trial run that will be discussed in more detail later. The feature C–C–N where both single bonds are ring bonds occurs in 0.177 (17.7%) of the file. Therefore, in a set of 33 compounds active in the trial system, 5.83 are expected to have this fragment, assuming it has nothing to do with activity. Moreover this number should follow the binomial distribution with a mean of 5.83 and a standard deviation of $[np(1 - p)]^{1/2}$ or 2.19, using $p = 0.177$ and $n = 33$. In our example of 33 actives, the actual number containing this feature was 11. Thus, the number of standard deviations (SD's) away from the mean is $(11 - 5.83)/2.19$ or 2.36.

This number, 2.36, could be used as the weight for this feature, but we can consider further the probability P that we can get 2.36 SD's away by chance. Using the normal approximation to the binomial distribution a statistical table shows that the two-tailed value for P is 0.0183. Since the weight should be inversely related to P (the smaller the P, the larger the significance), $1/P$ is used. And since the weights are to be added $\log 1/P$ is used, which is 1.75 in this case. P is thus a probability related to the fact that a compound is active, and $\log 1/P$ is a measure of this probability. $\log 1/P$ is used because it gives convenient magnitudes and has the following additivity properties.

Forming the score for a new compound by adding the weights for each feature corresponds to multiplying the probabilities, P_i, for each feature. That is

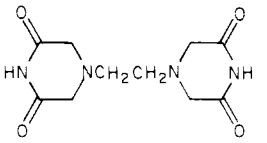
$$\Sigma(\log 1/P_i) = \log \Pi(1/P_i) = \log 1/\Pi P_i$$

Σ represents summation and Π represents multiplication, both over the index i. Thus, there is no assumption of additivity such as is found in the Free–Wilson method although the effect is similar. We are merely computing a measure of the probability of activity of a compound based on the statistical evidence of its features. Thus, the resulting score is not an estimate of the degree of activity but an estimate of the probability that the compound belongs to the active set. Multiplication is a proper way to combine probabilities under the assumption of independence of features. More will be said about this assumption later as it ties in with our discussion of redundancy among features.

In the above illustrative computation of the weight for a feature, the feature appeared more often in the active set than its expected value. If it had appeared less often then it would be considered to be a negative number of standard deviations from its normal value and its weight would be computed in the same way but with a negative sign.[10] Thus, the negative weights are additive in the same way as the positive weights and can be considered as a measure of the probability that the compound does not belong to the active set. Equal parts of negative and positive weights can cancel each other out.

The conversion from a number of standard deviations to $\log 1/P$ is plotted in Figure 1. Logarithms are taken

Table I.[a] Compound No. 129/942, Active Score = 20.535, Inactive Score = −3.587



Key no.	Key		Actual occur.	Activity wt	Inactivity wt	Wt occur.
17	MISC	TM (tautomer)	1	1.741	1.875	1
18	MISC	UN (universal)	1	0.000	0.000	1
30	RSIZE	6	2	−0.366	−5.146	1
30	RSIZE	6	2	−1.333	−1.901	2
95	RING	C 4 N 2 (0,2)	2	3.095	0.688	1
95	RING	C 4 N 2 (0,2)	2	1.558	−0.169	2
153	NUC	C 4 N 2 (0,2)	2	4.594	−0.326	1
153	NUC	C 4 N 2 (0,2)	2	0.978	0.000	2
258	AAKEY	C C N N1 N1	2	0.952	0.441	1
258	AAKEY	C C N N1 N1	2	0.594	1.498	2
272	AAKEY	C C O R1 NT	4	0.757	0.952	1
272	AAKEY	C C O R1 NT	4	1.831	−0.728	2
272	AAKEY	C C O R1 NT	4	2.724	0.000	3
319	AAKEY	C N O RT NT	4	1.325	0.955	1
319	AAKEY	C N O RT NT	4	0.345	−0.253	2
319	AAKEY	C N O RT NT	4	0.178	−0.529	3
340	AAKEY	N C C R1 N1	4	0.332	0.367	1
340	AAKEY	N C C R1 N1	4	−0.144	0.357	2
340	AAKEY	N C C R1 N1	4	1.375	−1.669	3

[a] The scoring of a sample compound from the study to be described later. Note the multiplicities of key occurrences, discussed in the section on eliminating redundancy.

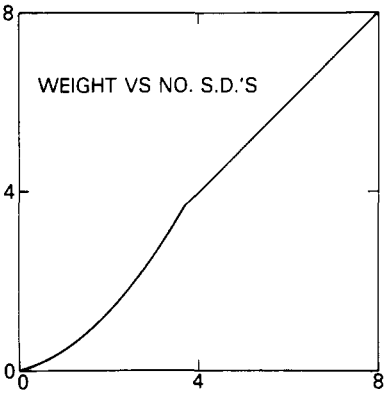


Figure 1. The weights are determined by the number of standard deviations according to the log $1/P$ relation, changing to the identity relation for larger numbers of standard deviation. P is the two-tailed statistical P value for the corresponding number of standard deviations.

to base 10. Statistical tables rapidly run out when the number of standard deviations exceeds 4, because P in the normal distribution decreases faster than exponentially. For these larger number of standard deviations, the number of standard deviations is used as the weight of the feature. The two weighting functions mesh at about no. of SD's = 3.6 as is shown in Figure 1.[11]

The weights and scores log $1/P$ that will be shown to be obtained from our data would give extremely low values in many cases if converted back to P values. That is because the active and inactive sets are compounds that were actually chosen purposely rather than randomly and may not be representative of the file. Under these conditions, and since there are no readily available assumptions about the selection of compounds for the training set, the use of log $1/P$ gives a good spread of values, while retaining the appropriate order.

Each compound receives a score which is the sum of the weights for all its features. As discussed earlier, the set of known active compounds provides an activity score for a given test system, and the set of inactive compounds provides an inactivity score. An example of a compound's keys, weights, and scores is shown in Table I.

Establishing Priorities. The activity and inactivity scores can be combined in various ways depending on the needs of the selection program and based on the results on the training sets. For example, if compounds were desired with structures unlike those that have been extensively tested, those compounds with low scores, especially low inactivity scores, would be selected. That is, a rule would be established under which compounds with low activity and low inactivity scores would be given higher priority for acquisition for testing than a compound with both a high activity and a high inactivity score.

A simplier rule would be the use of a threshold for activity and a threshold for inactivity, depending on the range of the training set scores. The thresholds can vary, depending on the number of false positives one is prepared to admit.

Another variation, which is used in the experiment described below, is to combine the activity and inactivity scores into a single value, to be used as a priority value for testing in the specified system. This variation recognizes that any statistical method is not foolproof and should not be the basis for the complete exclusion of a compound from all screening. Even the compounds assigned the lowest priority by a statistical method can be selected for testing in other systems and can be selected for testing in any system for any other reason.

One type of rule can assign priorities by computing a linear combination of the activity and inactivity scores. If the activity and inactivity scores are plotted along two axes in a plane, each compound appears as a point whose coordinates are the scores. Then any linear combination of the two scores can be represented geometrically as a projection of each point onto a line of a given direction, the priorty being measured along that line. See Figure 2 for an example of the Fisher direction[12] which was designed to provide optimum separation between the known actives and known inactives of the training sets when they are

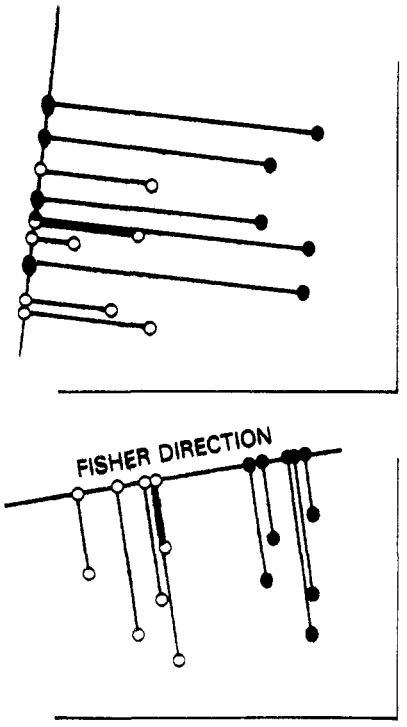


Figure 2. The projections of the dots and crosses onto two lines show that directions can be chosen to enhance separation of the projected classes of points. The Fisher direction is optimal under certain normality assumptions.

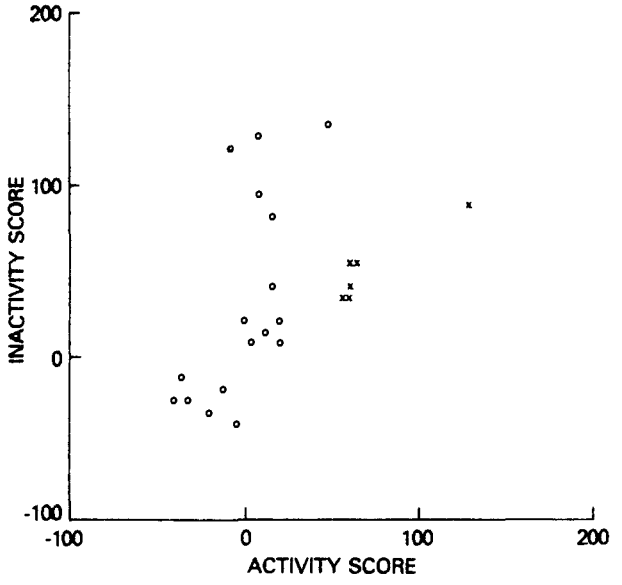


Figure 3. The activity scores (abcissa) and inactivity scores (ordinate) of the 24 compounds that were treated as unknowns in ref 12. The seven actives were labeled X. Two of the actives overlapped, so there are only six X's in the plot.

projected along this direction. This method was used in the study to follow. Establishing priorities allows the easy use of rank tests to provide a measure of effectiveness on the unknowns.

Mouse Ependymoblastoma Study. A set of 170 compounds, which had been tested in mouse ependymoblastoma,[13] was selected for an experimental trial. These compounds covered a broad range of structure classes and, moreover, have recently been the subject of a structure–activity methodology study by Chu et al.[1] involving prediction by nearest neighbor and learning machine methods. One of our objectives was to see whether our simpler method led to equally good results.

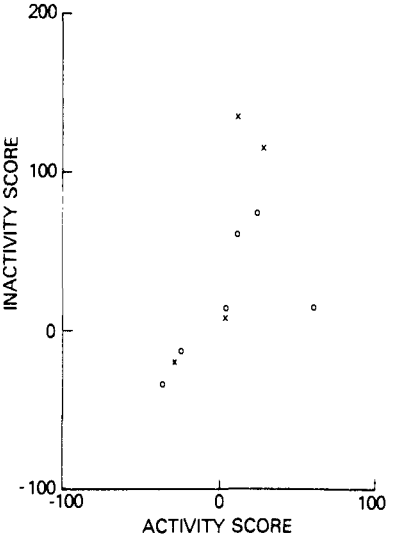


Figure 4. The ten additional unknowns in our data. These actives (X) are all marginal, with T/C of about 125%.

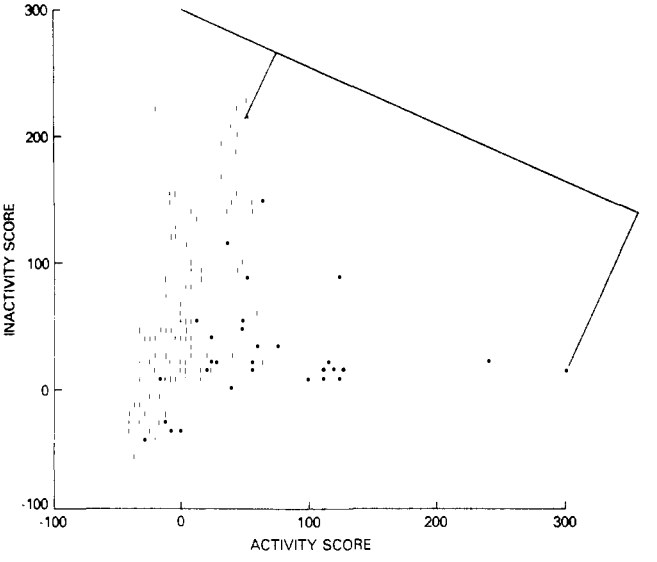


Figure 5. A plot of the known actives (A) and inactives (I) of the training set. The Fisher direction provides an idea of the kind of separation that can be expected from our method.

The training set we used was substantially the same as in Chu et al., consisting of 33 actives and 103 inactives. Prediction was done on precisely the same set of 24 compounds which were actually incompletely tested at the time Chu was predicting them. Seven of the 24 compounds subsequently turned out to be active. While the reliability of the biological data would be improved with the criterion for activity set at a T/C^{14} of 150%, a T/C of 125% was used in order to provide a larger number of active compounds. The resulting data were not completely satisfactory because six of the seven active unknowns were quite similar to each other in their structure features and scores based on them. Ten more unknowns, which had been tested later, were also predicted. These ten compounds had four actives, all in the marginal T/C range of 125–150%.

The resulting separation was excellent on the original set of 24 unknowns and compared favorably with the results of Chu et al. In Figure 3 we show the activity and inactivity scores of this set with the seven that were active represented as X's and the others as O's. The additional ten unknowns are shown in Figure 4, and we see that the resulting separation is not nearly as good.

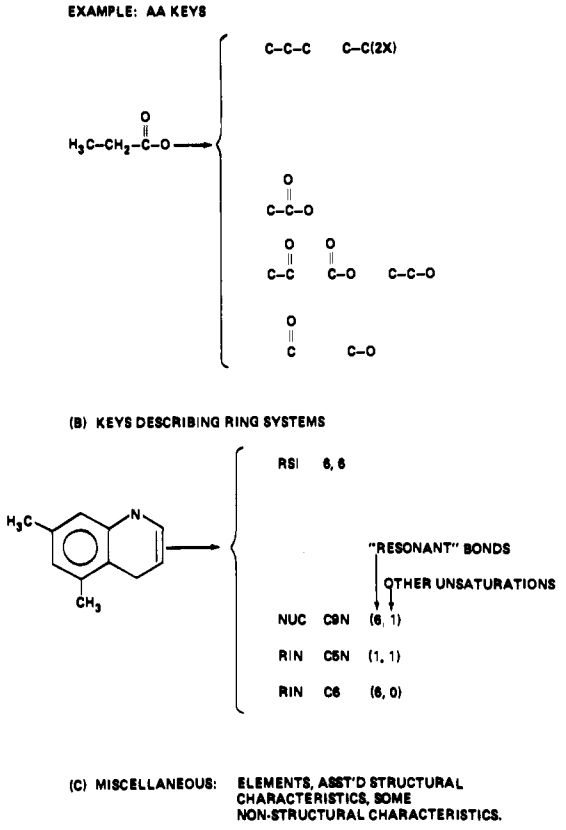


Figure 6. Illustration of keys on the NCI chemical file; courtesy of M. Milne and G. Hazard.

The training set itself, consisting of 136 compounds, can be plotted in the same way as is shown in Figure 5. The actives are represented by the letter A and the inactives by the letter I. Also a line in the Fisher direction is shown. When all the 34 "unknown" compounds, of which 11 were active, were arranged in the priority ordering determined by the Fisher direction, the sum of the ranks of the 11 actives was 133 for a significance level of 0.014 using the Wilcoxon rank sum test.[15] At that level of significance we can reject the hypothesis that there is no difference between the active and inactive unknowns.

Chemical Structure Features. For the presentation of further results it is necessary to review the chemical structure features that were taken from the DR&DP chemical file. These were keys that are generated routinely for information retrieval purposes. It was hoped that they would be useful for a selection process as well.

The structure keys were developed under contract by the University of Pennsylvania[16] for DR&DP and are basically of two types: small "augmented atom" (AA) fragments and ring keys. See Figure 6. The AA keys each consist of a nonterminal atom and one or more of the atoms bonded to it, up to a total of five atoms. The ring keys can be further divided into simple ring, nucleus, and ring size keys. There are more than 10 000 structure keys because of the extensive bonding information they contain. In the AA keys each bond is distinguished along two parameters according to whether it is ring or nonring and, second, whether it is single, double, triple, aromatic, tautomer, or delocalized. Likewise, in the simple ring and nucleus keys counts are kept of two categories, the first being aromatic and delocalized bonds and the second being the total of all other nonsingle bonds. This bond refinement forces a situation where most keys occur extremely rarely in the file; indeed, less than 10% of the keys occur in more than $^1/_{1000}$th of the file.

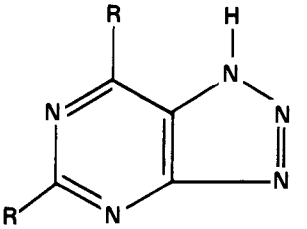


Figure 7. Structure in which redundant keys are difficult to recognize by merely examining keys.

Aside from being discriminatory for the purposes of retrieval, which these keys seem to be, structure features used for preselection must have sufficiently many keys among them that are relevant to activity in the various test systems. However, if there are too many irrelevant keys, there is a good chance that some of these may appear relevant in some of our data. This is the problem of noise, which is related to redundancy among keys.

Redundancy among Features. The problem of redundancy, or interdependence of features, is quite pertinent, if only because the method used in this study assumes independence of features. The net effect of redundancy is to cause some characteristics to receive more than their due weight, since they are represented in more than one key, and these keys are considered independently. Thus, allowing these redundancies leads to a further distortion on a system which is already imperfect, since many characteristics are not represented by complete keys but only by fragments which are sometimes connected, sometimes unrelated.

There are many obvious redundancies among the structure keys in the chemical file. For example, the benzene ring has the key RING C6 (6,0) which is present in approximately 66.6% of the compounds. However, there is also a nucleus key, NUC C6 (6,0), which occurs whenever a compound has a benzene ring which is not part of a larger ring system. This occurs in 53.6% of the compounds. Other dependent keys with these two are the C-C and C-C-C AA keys where the bonds are ring aromatic. These benzene-related keys happened to receive fairly large negative weights both for activity and inactivity. It was apparent that these keys contributed to poor results in two or three compounds in early trials. These results were improved by the simple elimination of redundancy described in the next section.

It can be seen from the foregoing example that the redundancies are not always simply related. Another example shows that they are not always easy to detect. The key RING C2N3 (1,3) seemed to occur together with the key NUC C4N5 (6,3) almost always since they have almost exactly the same incidence. This substructure is shown in Figure 7. On investigation it was found that the ring key occurred in 89 structures and the nucleus key in 86 structures, but their joint occurrence consisted of 70 structures. This shows a great deal of redundancy but not complete superfluity of either of the keys.

As is observed in Duda and Hart,[17] sometimes independence assumptions produce good results even though they are not justified. But because of this assumption of independence, redundancy should be eliminated wherever possible.

Eliminating Redundancies. A good example of key redundancy that was easily removable is that of keys which denote not merely presence but the multiplicity of occurrence of the structure fragment. Typically, the keys in the data did contain counts of the number of occurrences. It was decided not to use independent keys for the different multiplicities since a compound which has two

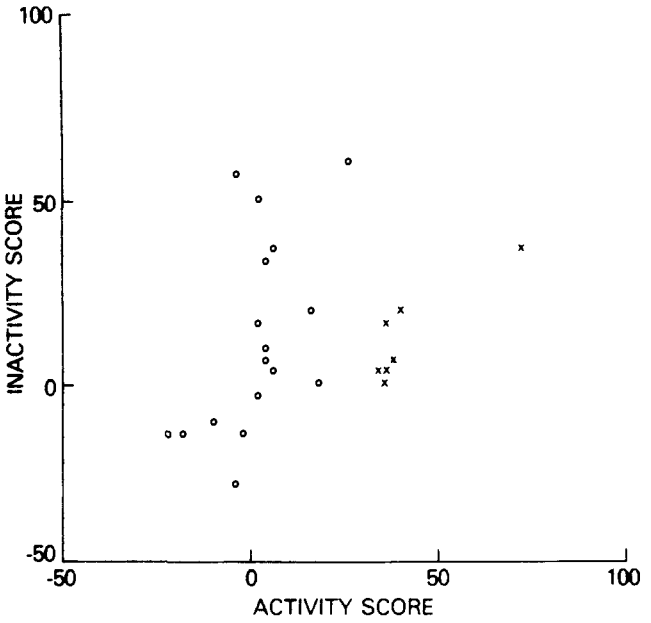


Figure 8. Compare with Figure 3, after some largely redundant features have been removed.

copies of a certain structure should also be considered to have one copy of this structure. Moreover, using each multiple as a new key produces keys which are highly interdependent. The only way to remove the redundancy was to use conditional probabilities, that is, not the actual probability that a compound has two (or more) occurrences but the probability that it has two (or more) occurrences given that it has one (or more) occurrence.

The expected number of active compounds having two (or more) occurrences of the structure can then be computed by multiplying the actual number of actives having at least one occurrence of the structure by this conditional probability of two (or more) occurrences given one (or more) occurrence. This conditional probability is easily computed from the frequency statistics by dividing the incidence of at least two occurrences of the structure by the incidence of at least one occurrence of the structure. In our experiment we considered multiplicities up to three (or more) occurrences.

As an experiment in eliminating redundancy among the keys, many of the AA keys were excluded as follows. Those which are atom pairs were eliminated because the information is contained in the three atom keys. Also excluded were the four atom AA keys since much of the information is already contained in the three atom AA keys and their counts. With this smaller set of keys, a new run of the training and priority assignment produced a rank sum of 117 for the 11 actives. This is significant at the 0.0027 level.

As a further experiment all AA keys with only ring bonds were excluded, as well as the even atom AA keys just described. These ring bond AA keys can be considered to be covered to some extent by the ring keys. The rank sum of the 11 actives was then down to 107, significant at the 0.0008 level. The results are shown in Figures 8–10.

Remarks and Conclusions

The experimental trial showed that on a small set of murine ependymoblastoma data this simple statistical-heuristic method using structural features as predictors of activity in a specific model compared favorably with a more sophisticated pattern recognition method. Simple experiments on removing redundancy among the features improved the results.

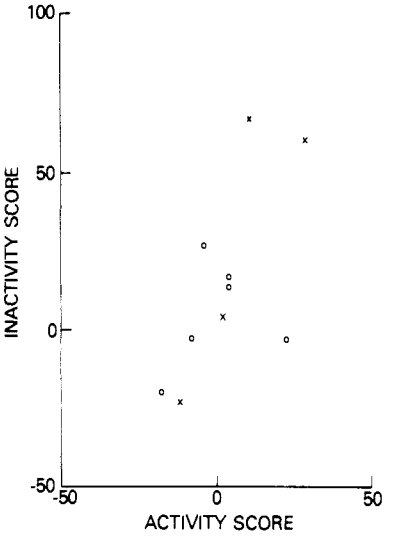


Figure 9. Compare with Figure 4, after some largely redundant features have been removed.

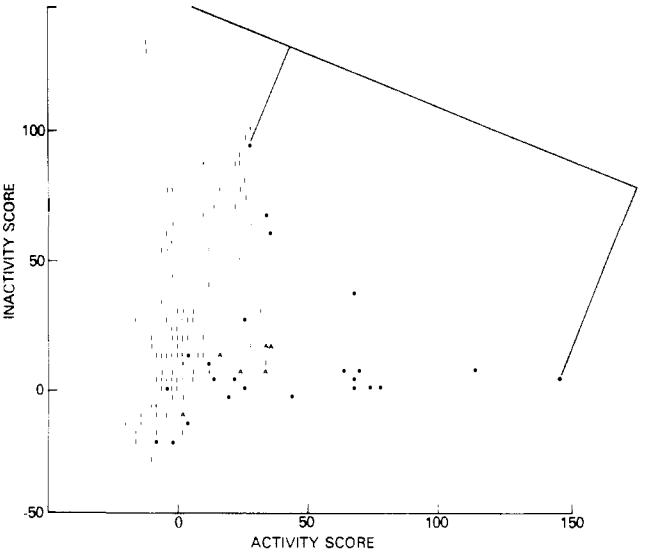


Figure 10. Compare with Figure 5, after some largely redundant features have been removed.

It will be necessary to verify in larger data sets this improvement in performance upon the simple exclusion of keys which are already somewhat covered by other keys. In any case, this is a fairly crude approach to removing redundancy.

A more exact approach would be to use the containment relations among keys as was used in the case of multiple occurrences. However, it is difficult to get direct inclusion conditional probabilities among the keys. Joint incidences are not available and most keys have more than one smaller key included, or possibly included, in their structures.

The main difficulty with the NCI structure keys lies in the finding of dependencies across different classes of keys. The AA key dependencies can be to some extent established among themselves. However, there is a big gap from the AA keys to the ring keys and indeterminacy in placement of the bonds in the ring keys makes direct dependence difficult to establish. Some of these difficulties can be appreciated by referring to the examples of interdependence cited earlier.

Thus, besides the extension of this work to more significant data, it will be productive to carry out procedures for eliminating redundancy among the features. Along this

vein, different sets of features will be tried. A completely different set of keys[18] designed for retrieval at the Walter Reed Army Institute of Research has the property that all the conditional probabilities are available from the generation of these keys in the form of a hierarchy according to information-theoretic principles.[19] It is also possible to generate features for the purpose of good biological discrimination, using the active and inactive training sets as guides.

The use of other feature systems must wait upon the full implementation of the new sets of features. Of course, to get a completely adequate set of features, we would need to include stereochemistry. One way to achieve three-dimensional features would be by autocorrelation on electron density maps. However, this is a fairly complex approach, which may be difficult to apply on a large scale.

The study of the feasibility of using this statistical-heuristic method for selecting drugs for testing in the P388 model is now under way and will be reported soon. Preliminary results show excellent separation of actives from inactives by means of the activity score alone. In the P388 data, it seems that including the inactivity score produces a small, perhaps marginal, improvement, and eliminating redundancy as by the elimination of keys produces a bigger, but still marginal, improvement. If further work on the P388 model bears out the early results, then it is possible that very sophisticated structure features may not be necessary, or merely marginal.

References and Notes

(1) K. C. Chu et al., *J. Med. Chem.*, **18**, 539 (1975).
(2) Entropy, Ltd., Pittsburgh, Pa., unpublished report, R. Christensen and T. Reichert, June 1975.
(3) W. T. Miller, *Cancer Chemother. Rep., Part 2*, **5**, 253 (1975).
(4) S. Richman and D. Lefkovitz, Abstracts of Papers, 169th National Meeting of the American Chemical Society, Philadelphia, Pa., April 1975.
(5) C. Hansch, *J. Med. Chem.*, **19**, 1 (1976).
(6) W. P. Purcell, G. E. Bass, and J. M. Clayton, "Strategy of Drug Design", Wiley, New York, N.Y., 1973.
(7) R. D. Cramer III, G. Redl, and C. E. Berkoff, *J. Med. Chem.*, 17, 533 (1974).
(8) Reference 6, p 89.
(9) The weights cited appear in the summary of ref 7 in G. Redl, R. D. Cramer III, and C. E. Berkoff, *Chem. Soc. Rev.*, 4, 273 (1974). The probabilities, however, are not calculated from the entire file but from the tested compounds. Their formula has the general effect of giving each feature a weight proportional to its incidence, for the same relative activity. The weights used in ref 7 are indeed different and are simply the number active divided by the number tested for each feature. This can give low incidence features an exaggerated effect. Our method falls between these two extremes. Statistical considerations yield weights that vary with the square root of the incidence, for the same relative activity.
(10) Among those structure features getting negative weight should be those which do not appear at all in the training set. If they are among the large number of low incidence keys then their expected number of occurrences would be very close to zero, anyway, so their weight, through negative, would be extremely small. We have, however, ignored all keys which did not occur in the training set. In a future version of the program weights will be computed for keys which do not occur if they have some significant expected number of occurrences, say 0.5.
(11) A reviewer has pointed out that, from Figure 1, it would be just as well to use the number of standard deviations as a weight, instead of the two-tiered system. This turns out to be true, even in a test on the P388 data. One can simply use the number of standard deviations as a weight and consider additivity as an heuristic. However, the theoretical rationale lies with the log $1/P$ weight. The heuristic of the number of standard deviations was forced on the principle author because keys 20–40 SD's from their expected value would get extremely large weights. When characteristics are repeated in several keys, those with large weights (in terms of standard deviations) increase their effect on the score toward the log $1/P$ direction.
(12) R. O. Duda and P. E. Hart, "Pattern Classification and Scene Analysis", Wiley-Interscience, New York, N.Y., 1973, p 114.
(13) R. I. Geran et al., *Cancer Chemother. Rep., Part 2*, 4, 53 (1974).
(14) T/C in this model is the ratio of the median life span of the treated animals (T) over the median life span of the controls (C), converted to a percentage.
(15) E. L. Lehmann, "Nonparametrics: Statistical Methods Based on Ranks", Holden-Day, San Francisco, Calif., 1975.
(16) M. Milne and G. F. Hazard, Abstracts of Papers, 169th National Meeting of the American Chemical Society, Philadelphia, Pa., April 1975.
(17) Reference 12, p 68. See also note 11, last sentence.
(18) A. P. Feldman and L. Hodes, *J. Chem. Inf. Comput. Sci.*, **15**, 147 (1975).
(19) L. Hodes, *J. Chem. Inf. Comput. Sci.*, **16**, 88 (1976).

Anticoccidial Derivatives of 6-Azauracil. 1. Enhancement of Activity by Benzylation of Nitrogen-1. Observations on the Design of Nucleotide Analogues in Chemotherapy

Banavara L. Mylari, Max W. Miller,* Harold L. Howes, Jr., Sanford K. Figdor, John E. Lynch, and Richard C. Koch

Medical Research Laboratories, Pfizer Inc., Groton, Connecticut 06340.

Received August 2, 1976

Benzylation of 6-azauracil at N-1 (which corresponds to the point of attachment of the ribose phosphate unit in pyrimidine nucleotides) has been found to augment its anticoccidial activity fourfold. The high potency of 1-benzyl-6-azauracil is ascribed to a combination of intrinsic activity, efficient oral absorption, and a moderate rate of excretion. Metabolism experiments using 1-benzyl-6-azauracil labeled with ^{14}C in the heterocycle and (separately) in the side chain showed that, in the drug accounted for, no cleavage had occurred. Additional activity increases were achieved by introducing small, electron-withdrawing substituents in the meta and/or para position(s) of the benzyl group. One of the most active derivatives, 1-(3-cyanobenzyl)-6-azauracil, is about 16 times as potent as 6-azauracil.

The exigencies of the world food supply have led to ever more intensive agriculture and animal husbandry. Large-scale enclosed poultry rearing has been made possible during the last 25 years by the discovery of feed-incorporated anticoccidials to control the most troublesome social disease of fowl. Coccidiosis is caused