Comparison of Batch and Time-Sharing Computer Runs for Correlating Structures and Bioactivity by the Hansch Method*

PAUL N. CRAIG**
Smith Kline and French Laboratories, Phila., Pa. 19101

Received May 19, 1971

The Hansch program for multiple parameter analysis of structure-activity relationships has been adapted from a batch computer format to a time-shared console operation. Turn around time has been greatly reduced, as have total elapsed time and relative costs. However, it was necessary to sacrifice some of the capabilities of the Hansch program for it to run on a terminal, so a trade-off is necessary. For a new series of data, a preliminary study is run on the terminal, and if results suggest reasonable correlations, these are followed later with full-scale computer runs by the batch process. This approach uses the best features of both systems and results in economies of both time and money.

Since 1964, the particular method of relating biological activity to chemical properties which involves the use of regression analysis to test for correlations between various combinations of partition, polar, and steric factors has become known as the "Hansch" method.[1] Hansch and colleagues at Pomona College have developed an elegant batch computer program for obtaining regression analysis results with data of varying complexity. It has provision for up to 24 independent variables, any one or combinations of which can be called for by means of a "weight" input card. One strength of this program lies in its legible output, which is readily understood by the chemist.

A rapid scan of the summaries is followed up by examination of the more complete data for an individual regression analysis. The output displays the original data input, together with the observed biological activity and the activity calculated by the obtained mathematical expression. The deviation between observed and calculated activities, as well as various statistical parameters from the regression, are part of the output. A description of each molecule is also given.

At this point one might say all is under excellent control, and why interfere with such a smooth operation? However, in carrying out a study of a series of compounds with interesting biological activity, it is rare that one obtains the ultimate degree of correlation desired, on the first try. Since the program is written for batch processing, this means that any changes which may be indicated as a result of a preliminary run, after proper adjustment of, or changes in, the input data, must be followed by a second, third, or fourth computer run. This is fine if you have the situation with which Hansch is favored—namely, a computer which is just a few hundred feet away from his office, and one which is nearly always available to him within a half-hour notice. In the industrial world which I inhabit, the corporate computer is operated in a different part of our corporate structure and is not used solely for research purposes. The result is that I can count on but one turn around in a day of operation. For some of the more complicated Hansch analyses which I have been carrying out in the past several years, it has been necessary to make up to 10 successive runs where each one will depend upon changes or results obtained from previous runs. With the five-day work week, this means an average of two weeks elapsed time from the time I first send data to our corporate computer to the time when I finally receive the last version. In actual practice, a much worse problem results than merely the one- to two-week time delay. I find it difficult to turn on the high level of concentration required to study these results for five minutes every morning at 9:00; instead, it requires 5 or 10 minutes for me to review the previous runs to see what to look for in the analysis underway. This is duplicated effort for each turn around.

About two years ago, I obtained a government contract to analyze the antimalarial drugs by means of the Hansch method for the Walter Reed Army Institute of Research. Since I could foresee a greater need for rapid interaction than before, I obtained a time-shared computer console in the IBM Call 360 Basic Time Sharing System. The particular system was chosen because of the IBM 2741 computer console, which is basically an IBM electric typewriter with an interchangeable typing element. This is connected to the computer by means of a dataphone. The fixed rental fee for the computer 2741 terminal is $100.50 per month. The dataphone rents for about $20 per month, and, in addition, a $12 charge is made for a private telephone line between my office and the central IBM computer, which is located a mile away in downtown Philadelphia. The 2741 terminal, which looks and handles like an IBM Selectric typewriter, types at a top speed of approximately 15 characters per second. This speed is achieved, of course, only on playback of a large set of data.

The Philadelphia system serves northern New Jersey, Philadelphia, Delaware, Baltimore, and Washington, D. C. Some who use this system in northern New Jersey and Washington have told me that there are sometimes delays

*Presented before the Symposium on "Input-Output Interaction of the Chemist with the Computer," Division of Chemical Literature, 161st Meeting, ACS, Los Angeles, Calif., March 31, 1971.

**Present address, Craig Chemical Consulting Services, Inc., 120 Stout Rd., Ambler, Pa. 19002.

```
1   Input X,Y
2   IF X ≤ 1 GO TO 10
4   LET Z = X
5   GO TO 100
10  INPUT A,B,D,E,F,G,H,I
20  LET Z = 12.01*A+1.0078*B+14.01*D+16*E+34.457*F+19*G+79.91*H+32.064*1
100 LET M = .43429*LOG(1000*Z/Y)
110 PRINT USING 120, Z,Y,M,
120 WHEN MWT = ###.## AND ED50 = ###.###MG, LOG 1/C = ##.###
125 GO TO 1
130 END
```

Figure 1.

involved with being located a greater distance from the computer. These are usually involved with problems of sharing a telephone line. My terminal was hooked up to a private line, and also was only about a mile away from the main computer. I rarely have had to wait to get on the system, but there have been infrequent but annoying breakdowns of the central system, resulting in the loss of whatever data one was working on at the time. This happened about four times in an 18-month period. One quickly learns to store important data at regular intervals as insurance against this type of loss.

For each hour that the terminal is connected to the central computer, there was an $11 per hour rental charge. For each minute of central processing unit time (CPU) used there was an additional charge of $7. My work typically averaged about 1 minute of CPU time per hour at the terminal; hence, the charges were about $18 per hour in addition to the fixed fees of $132 per month. Over a one-year period my total computer costs were about $4500, for the use of this system for about 150 hours of connect time. There are additional charges for unusual storage requirements, but for this type of work, these are minimal. One must be careful not to save duplicate copies of lengthy programs or data over long periods, or the storage charges can mount appreciably. The standard storage space which is included in the above pricing proved to be more than adequate to store several regression-type programs without extra charge.

Because of the limitations of this particular system, I was unable to place the entire Hansch program onto the system. This is basically due to the limitations on matrix size for this system, although other systems vary in this respect. The IBM 360/40 computer used by Hansch easily handles a matrix of 25 variables and could handle many more. In the Call 360 time-sharing system, there is a core limitation on matrix size which limits one to 15 variables.

Hansch has exceeded 125 compounds in a single run. The time-sharing computer also has no limitation in the length dimension of the matrix. For practical purposes, I have had to limit the number of dependent variables to 3 for the time-sharing computer. This is because of the expense of operating the computer for this type of regression analysis. A simple matrix inversion program such as used on the IBM 360 STATPACK programs for multiple regression analysis operates with perhaps one to two seconds CPU time for a medium sized matrix. Because of the specific requirements of the Hansch program, which result in a much more meaningful output than the other system, the amount of time for CPU time for such a run is multiplied by four- to fivefold. This quickly can add up to increased cost; consequently, I have used the limitation of 3 dependent variables at one time. However, this is adequate for most of the analyses studied.

The preparation of data for imput either to the Hansch batch method or to the time-sharing computer is a time-consuming operation. It requires careful preparation of data and key punching for the batch processing; the same preparation of data is followed by typewriter input on the time-shared console. Here another use of time sharing comes to the fore. Simple programs can aid materially in the preparation of the original data for analysis by these multiple parameter techniques. Such a program is shown in Figure 1. This program takes the data from the biological test results expressed in mg./Kg., and the molecular weight of the compound, or the molecular formula if the molecular weight is not available, as input, and converts this to the log 1/c value (the log of the reciprocal of the molar concentration) which gave a predetermined biological effect. The use of this simple program has speeded work and improved accuracy greatly. The inverse of this program has also been written. After an analysis has been carried out and when predictions of new compounds have been made, this allows one to convert the prediction from a log 1/c value back to mg./Kg. of the proposed compound.

The summary for a typical run is very much like that of the batch computer run, although I have added some more data than Hansch usually uses in his summary to reduce the times when I would have to request a full printout of complete details. The Hansch program has been revised so that the first output is now the summary; then I can judge from the summary which analyses to see in greater detail. This is an important provision, since I am printbound with a relatively slow speed typewriter of 15 characters per second as opposed to the high speed computer printout of 800 to 1000 lines per second. This particular run, involving some 30 compounds, required some 10 or 12 minutes to type in the input data (it would have taken me longer than that to keypunch the data for a batch run, although a skilled keypunch operator could do it in 5 minutes). After typing, it took the computer two seconds of CPU time and about 10 seconds elapsed time before the results began to be returned. After 3 minutes the summary of eight regressions had been completely printed out, and I was able now to judge which of the combinations of parameters I wished to study in detail. It required approximately 3 minutes to type out each desired complete set of details analogous to those which would result from a computer batch run. I now could make a careful study, select compounds for which data should be changed, or remove them from consideration. Then, on typing new instructions, in a matter of a few minutes the whole process could be repeated. In this way, one can quickly carry out a complete iterative process, and in one 2½- to 3-hour period on the computer, I carried out 6 successive changes and completed a complex analysis from start to finish. These three hours on the computer contrast with what would have required 6 to 8 days on the batch computer system previously described. The various tradeoffs that are required in analyzing the two systems are shown in Figure 2. The greatest saving, of course, is in over-all elapsed time from the beginning of an analysis to the completion. This is obtained, however, at the expense of an increased amount of the chemist's time required for the analyses. This is mostly taken up in output time re-

	PREPARATION OF DATA	INPUT	OUTPUT	TURNAROUND TIME
BATCH	SAME TIME REQUIRED FOR BOTH METHODS.	KEYPUNCHER—5-10 MINS.	HIGH SPEED PRINTER 1000 LINES/MIN.	UP TO 24 HRS.
TIME SHARING		CHEMIST—10-12 MINS. AT 2741 KEYBOARD.	15 CHARACTERS PER SECOND = MUCH SLOWER.	10-20 MINS. OR LESS.

Figure 2.

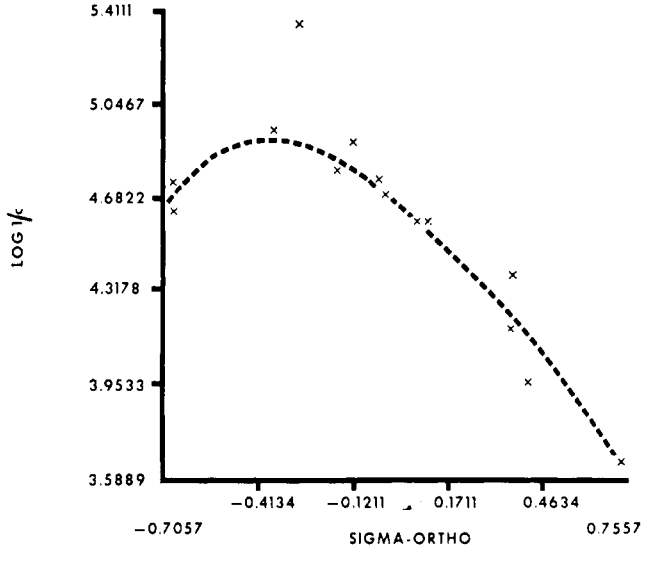

Figure 3. Reproduced from *Journal of Medicinal Chemistry*[2]

quired for typing back data on the relatively slow typewriter available in this particular set-up.

We have recently obtained a System 3/Model 6 IBM computer, and I have had the same Hansch program adapted to it. The same 3-hour analysis could be handled completely in less than 1½ hours using this particular system. Great savings arise from the much faster print speed, which approaches 60 characters per second. In addition, the System 3 computer has a cathode ray tube visual display unit, which is very helpful. This is a completely self-contained third generation computer with a small memory, but with two disc storage units. It takes from 6 to 10 minutes for a program of the complexity of the one just illustrated for the time-shared computer to run on the System 3 computer, as opposed to the 10 seconds or so elapsed time and only 2 seconds CPU time for the time-sharing computer. However, this is quickly made up in the output printing stage, since the new System 3 computer takes approximately 7½ minutes to print data for which the time-shared 2741 terminal required 30 minutes.

In practice, I now utilize the time-shared computer for the first study of a new series of compounds. Either the first attempt at analysis is a successful one, or leads to several rapid interations which result in a good analysis, or else a more complicated analysis is indicated. In that case, I switch to the batch computer method whereby more complex analyses can be run, albeit more slowly. An overall summary of pros and cons for both systems would have to list the following points. The batch computer has the greatest power, the greatest size and memory, the most rapid speed for CPU time required for such an analysis. However, in the organizational set-up so commonly found in industry, it is impractical to expect to obtain more than one computer run per day or even every other day using the large batch computer. On the other hand, the time-shared computer has a more limited capability in terms of the size of the matrix which can be employed, and requires more CPU time for operation than the batch computer, but has the great advantage of giving results while one waits. Thus, the interactive or iterative process of development of an analysis is greatly speeded up by means of the time-shared computer. Finally, although we have just begun to use the System 3 computer, it is obvious at this stage that the newer computers can play a very useful role. Especially, this will be true when some programming changes are made so that the cathode ray tube can be used, for example, in scanning the summary or selected items of output.

An unexpected bonus for use of the remote console was the ability to use the same set of input data that were used for the regression analysis with the STATPACK Statistical Programs which are a part of the Call 360 system available to all subscribers. Among these programs are several which enable one to obtain two-dimensional graphical plots. Thus, one can generate a plot from the original data and compare that with the equations obtained by regression analysis. This is useful only when a linear or parabolic relationship is being obtained. An example is shown in Figure 3.[2] A nice improvement was made last year when the graphical plotting capabilities were augmented by increasing the resolution obtainable on the 2741 terminal by using a separate typing sphere which had the ability to print 30 characters per inch instead of the usual 10 per inch. Other STATPACK programs of value to one working in this field are the multiple regression, correlation, and stepwise regression programs. These supplement the Hansch program and can be useful adjuncts.

LITERATURE CITED

(1) Hansch, Corwin, *Accounts Chem. Res.* 2, 232 (1969).
(2) Craig, P. N., H. C. Caldwell, and W. G. Groves, *J. Med. Chem.* 13, 1079 (1970).