~~My letter~~?

I suggest minimizing semantic argument.

D. Smith
B. Buchanan &
3/11/74
J. Feigenbaum (if you wish)

A note has recently appeared in this journal[1] entitled "Artificial Intelligence". This note cogently and aptly criticized an earlier paper[2] which used the same phrase in its title[3]. ~~Because we agree with this criticism and because the term artificial intelligence is being used among chemists more and more frequently, we wish~~ However, we believe it is important to distinguish model-free statistical pattern recognition schemes from other techniques used in artificial intelligence programs.[4]

~~Various statistical techniques subsumed by the term "pattern recognition" have been considered a sub-division of artificial intelligence in the past[4]~~. The procedures for pattern recognition developed in the early 1960's were far less purely statistical and more oriented to semantic information processing than ~~the pattern recognition work of~~ has prevailed in the 1970's. More recently, ~~however, the areas have~~ heuristic problem-solving has diverged ~~significantly~~ further[5] because of fundamental differences in initial assumptions and computational procedures. Although there is still no precise definition of artificial intelligence (AI), most workers in the area would agree that work in AI is characterized by its use of judgmental rules for reasoning about a problem. The judgmental rules, or heuristics, ~~certainly~~ do not guarantee the solution to a problem. ~~However, the heuristics do~~ they are intended to keep the scope of the reasoning steps of the program within bounds of plausibility. That is, an AI program may not solve a problem, but its reasoning, if the judgmental rules are good, will not be considered totally irrelevant. A second dimension which may help to distinguish AI programs from others is that the problems which are of interest are by definition ~~generally the kinds of problems which are~~ more complex than a ~~person~~ contemporary expert knows how to solve using a straightforward, algorithmic method. ~~Very~~ typically, the problems are non-numerical, that is, they are not the kinds of problems one can solve with a set of simultaneous equations.*

Computer programs with some degree of AI content are now being applied to chemical problems, for example, the analysis[6] and synthesis[7] of molecular structures. Even a cursory comparison of these reports with descriptions of applications of statistical procedures to chemical problems[8,9] will reveal fundamental differences in methodology.

The statistical procedures, described variously as, for example, machine intelligence[8] and pattern recognition[9], can be valuable techniques if applied with discretion. The fundamental assumption is that there is some relationship between the experimental data and the property (i.e., pharmacological activity[2]) of interest[8,10]. If this assumption is not correct, erroneous hypotheses may appear to be validated because of accidental clustering, as Clerc et.al.[1] and Perrin[3a] have shown.

~~In a sense~~, these statistical techniques can be called "~~mindless~~". model-free; in the sense that Judgmental knowledge used routinely by chemists is not employed ~~by these techniques.~~ As long as theoretical reasons for clustering‡ are lacking, interpretation of the results will be on a questionable footing. This is in sharp contrast to current AI programs where the judgmental knowledge and the ~~program's~~ reasoning steps are ~~well understood.~~ integral to the program's design.

[*except in the sense that all problems can be expressed as optimizations of Boolean matrices]

‡ see addendum.

The assignment of the correct number of degrees of freedom poses some of the subtlest problems in statistical analysis. For example, consider the series of chemical names for the alkanes:

odd	even
methane	ethane
propane	butane
pentane	hexane
heptane	octane
~~nonane~~	~~decane~~

It will be noted that for these first eight examples, there prevails a perfect agreement between the parity of the name and of the molecular formula. The statistical significance of this correlation is not to be defended, and its material and historical basis, if any, is a matter of linguistic rather than chemical theory. This "clustering" may, however, well be transmitted to thousands of derivatives whose names may then exhibit highly significant correlations with other properties.

This may seem a trivial level of correction. However, more generally, one must keep in mind that the sample of compounds on which characteristic data are available are always highly selected to start with, and that conventional statistical methods may be unable to remove the variety of sources of confounding. On the other hand pattern analysis may be a valuable approach to the furthering of speculations about functional signatures, which can then be subjected to further study for their possible theoretical significance.

References

1. J. T. Clerc, P. Naegeli, and J. Seibl, *Chimia*, **27**.

2. K. H. Ting, R. C. T. Lee, G. W. A. Milne, M. Shapiro, and A. M. Guarino, *Science*, **180** (1973) 417.

3. For additional criticism and a rebuttal, see a) C. L. Perrin, *Science*, **183** (1974) 551; b) K. H. Ting, *Science*, **183** (1974) 552.

4. E. A. Feigenbaum and J. Feldman, eds. "Computers and Thought" McGraw-Hill, New York, 1963, pp. 235 ff.

5. ? Newell

6. D. H. Smith, B. G. Buchanan, R. S. Engelmore, A. M. Duffield, A. Yeo, E. A. Feigenbaum, J. Lederberg, and C. Djerassi, *J. Amer. Chem. Soc.*, **94** (1972) 5962.

7. a) E. J. Corey and W. T. Wipke, *Science*, **166** (1969) 178; b) E. J. Corey, W. T. Wipke, R. D. Cramer III, and W. T. Howe, *J. Amer. Chem. Soc.*, **94** (1972) 421; c) H. Gelernter, N. S. Sridharan, A. J. Hart, S. C. Yen, F. W. Fowler, and J. J. Shue, *Fortschr. Chem. Forsch.*, **41** (1973) 113.

8. T. L. Isenhour and P. C. Jurs, *Anal. Chem.*, **43** (1971) 20A (August).

9. B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.*, **94** (1972) 5632.

10. B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.*, **96** (1974) 916.