A Note on Measures of Screening Effectiveness in Chemical Substructure Searching

DAVID BAWDEN* and JEREMY D. FISHER
Pfizer Central Research, Sandwich, Kent, CT13 9NJ England

Received April 24, 1984

A novel measure of the effectiveness of screen set performance for chemical substructure searching is described. It has advantages over the generally used screen-out measure, particularly for evaluating the performance of operational systems, and is unaffected by changes of file size and composition. It has been applied in screen-set design and is suitable for long-term monitoring of system performance and for comparison of screening performance in different searching systems.

INTRODUCTION

Considerable effort has been devoted to the derivation of measures of retrieval effectiveness for computerized information systems,[1] and a variety of such measures have been used in experimental systems.[2-4]

These measures have been designed primarily for document retrieval systems, with each item described by index terms, on the basis of which items are retrieved. With each item designated a priori as relevant or irrelevant to a query, the whole set of documents could be partitioned into relevant and retrieved, relevant and not retrieved, irrelevant and retrieved, and irrelevant and not retrieved. From these data, measures of effectiveness are calculated.

The two such measures to have gained widespread use are recall (the proportion of relevant items that are retrieved) and precision (the proportion of items retrieved that are relevant). Cleverdon points out that these measures represent "fundamental requirements of users, and it is quite unrealistic to try to measure how effectively a system or subsystem is operating without bringing in recall and precision".[5]

The use of measures of this sort in quantifying the effectiveness of computer-based chemical structure searching systems is not straightforward. In the simplest of such systems, with retrieval based solely on fragment code (a form of indexing exactly analogous to that for document retrieval), recall, precision, and similar measures are entirely applicable. For systems employing full structural representations, with an atom-by-atom search routine, however, recall and precision (as generally understood) must be 100%, since there can be no ambiguity in the matching of query substructure with the structures in the file. It is of course possible that the query input facilities (graphics, notations, etc.) will not permit certain forms of query (usually very general ones), but this cannot be quantitifed as a failing in effectiveness of the system.

Since, however, virtually all chemical information systems have multistage searching (most commonly screening, followed by atom-by-atom searching), there is a need for measures of the effectiveness of each stage of the search. In particular, it is necessary to quantify the effectiveness of screening, the rapid matching of relatively simple characteristics (fragments) between the query substructure and each structure in the file, so as to rapidly minimize the number of structures that must be checked by atom-by-atom searching, a highly computer-intensive process.[6] Such measures are essential for the choice of the most appropriate screen sets for particular chemical structure files.[7] The work reported below stems from our experiences in the design of an optimal screen set for the Pfizer chemical information system (SOCRATES). This is a connection table based system, with retrieval by fragment screening and atom-by-atom matching, following graphical query input.

SCREENING EFFECTIVENESS MEASURES

It is usual to assess the performance of the screening phase of a chemical substructure search by a simple measure of screen-out, i.e., the percentage of the file not passed to atom-by-atom search.[8] It is taken for granted that the higher the screen-out, the better the screen-set is performing. The implicit assumption is that the number of structures found to match at the atom-by-atom stage will be negligibly small. It is also assumed that alternative screen-sets that may be used do not differ markedly in efficiency of computer implementation.

There are, however, cases where this does not seem a sensible measure. Consider as examples two searches. In the first, 2% of the file is passed to atom-by-atom matching (98% screen-out), and all these structures are hits. In the second, 0.5% of the file is passed (99.5% screen-out), and only one out of five of those passed are hits at the atom-by-atom stage. Since the screens in the first case are operating optimally, it does not seem sensible to regard the screening as inferior to that in the second case, which the screen-out measure would imply.

Such factors may not pose problems for formal evaluation of screen-set performance in a experimental situation, based on test queries. With an operational system, however, the type of query is inherently unpredictable; particularly, end-user searching is the norm. Also, the files being searched in this case will be constantly altering in size and composition. There is therefore a need for a single informative and robust measure to deal with these factors, which can be used for initial screen-set design and for subsequent monitoring of system performance.

E_S MEASURE

The sole raison d'etre of a screening system is, of course, to reduce, so far as possible, the extent of atom-by-atom searching that must be undertaken. The most appropriate measure of screening effectiveness is therefore the number of *unnecessary* atom-by-atom searches carried out, i.e., the number of structures passed as potential bits at the screening stage and subsequently found not to match the query. This has an apparent similarity with the "fall-out" measure (number of nonrelevant and retrieved/total number of nonrelevant)[2,4] but is in fact a composite measure of effectiveness, involving both the screen-out and the precision measure for the atom-by-atom search. It is a valid measure for comparing performance of different screening sets for the same query on the same file of structures. However, it is more generally useful to have a measure that is applicable to different files, or to a single file growing with time, and to different queries. It is in fact easy to generalize this measure of screening effectiveness.

Let S be screen-out (real number between 0 and 1) and P be precision (real number between 0 and 1) where

$$S = \frac{\text{number of structures rejected by screens}}{\text{total number in the file}}$$

and

$$P = \frac{\text{number of hits from atom by atom}}{\text{number of structures passed by screens}}$$

0095-2338/85/1625-0036$01.50/0

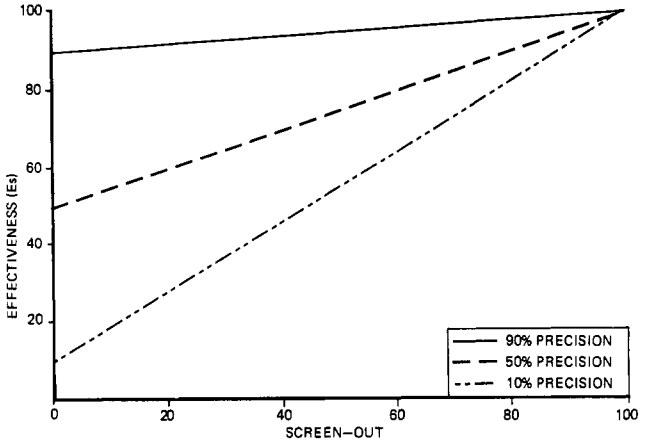

Figure 1. E_s measure against screen-out: eq 1.

Let T be the total number of structures in the file and f be the number of structures passed to atom by atom that are *not* hits. Then

$$f = T(1 - S)(1 - P)$$

To make the measure independent of file size, and to express it as a percentage

$$F = 100(1 - S)(1 - P)$$

Then, the corresponding measure of screening effectiveness, E_s, will be

$$E_s = 100[1 - (1 - S)(1 - P)] \quad (1)$$

This gives an intuitively sensible measure. E_s is 0, indicating total failure, if $S = 0$ *and* $P = 0$; i.e., the whole file is passed by the screens, but there are no hits. E_s is 100, indicating total success, if $S = 1$ *or* $P = 1$; i.e., either the whole file is rejected at the screening stage, or all the structures passed (regardless of the screen-out value) are hits.

The E_s measure is governed by P if $1 - S$ tends to 1: i.e., S tends to 0, as would be intuitively expected. It is governed by S if P tends to 0. This in turn can result from the number of structures passed by the screens becoming very large (i.e., total failure of screening) or from the number of hits becoming very small.

A plot of E_s against S for three values of P is shown in Figure 1. The main drawback can be clearly seen. In the practical operation of any realistic screening system, anything other than an absolutely disastrous search will give an E_s value in the high 90s; i.e., the measure is insensitive, under the most likely circumstances of its practical use.

The simplest way of correcting this is to define a corrected E_s measure, E_s', as

$$E_s' = \frac{100}{100 - N}(E_s - N) \quad (2)$$

where N is the lowest value of E_s to be taken as representing a worthwhile screening performance. If N is 90, then

$$E_s' = 10(E_s - 90)$$

thus effectively extending the range of E_s' from 0 to 100 to −900 to 100. All negative values are taken as 0, i.e., equivalent to total failure. A graph of E_s' against S for the same value of P is given in Figure 2.

Alternatively, a continuous transformation may be applied, so as to avoid the "cut-off" in the E_s' measure. This has the effect of equivalencing a large value of $E_s(v)$ to a smaller value

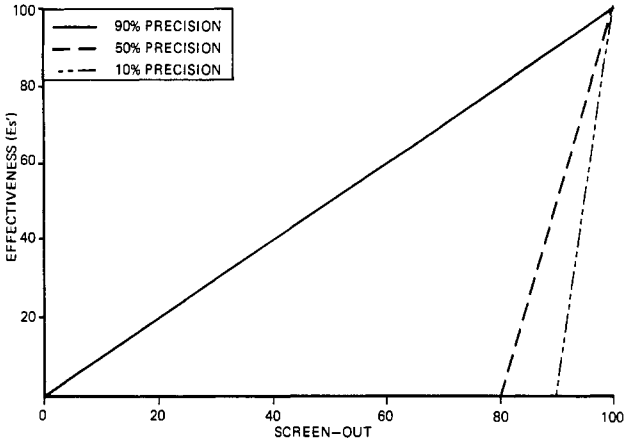

Figure 2. E_s' measure against screen-out: eq 2.

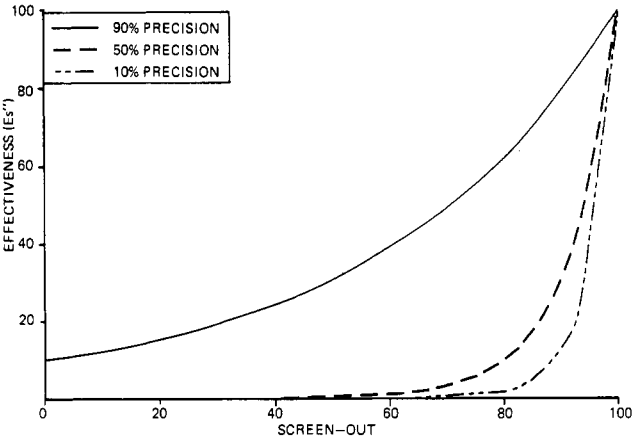

Figure 3. E_s'' measure against screen-out: eq 3.

of $E_s''(v_t)$ and "stretching" the upper range of E_s accordingly.

A suitable transformation is

$$E_s'' = E_s k B^E$$

where

$$k = (v/v_t)^{-(100-v)}$$

and

$$B = (100 - v)(v/v_t)^{1/2}$$

Thus, if for example the 90–100 range of E_s is to be "stretched" to cover the 10–100 range of E_s'' ($v = 90$, $v_t = 10$), then

$$E_s'' = E_s \times 9^{(E_s/10)-10} \quad (3)$$

A plot of E_s'' with these values is shown in Figure 3, for the same three values of precision as in Figures 1 and 2. The choice of the most appropriate variant of the screening effectiveness measure is a matter of experience in use.

PRACTICAL APPLICATION

The E_s measure, displayed in conjunction with screen-out, has proved to be a very useful measure in the development of SOCRATES, in assessing success of screening, in choosing between alternative screen-sets, and in identifying particular sorts of queries for which augmentation of the system is desirable. In the longer term, automatic logging of the E_s measure for each search carried out will enable accurate monitoring of the success with which the system is being used and will identify any problem areas for which the screen-set requires modification. Both measures are, of course, unaffected by changes in file size and composition. This effectiveness measure may also be used for informative comparisons

of searching quality on different substructure searching systems, though with the proviso that care must be taken to ensure that differences in graphical, or other, query input language do not invalidate the comparisons.

CONCLUSIONS

The effectiveness measure described above is a novel approach to the quantitative assessment of screen-set performance. It overcomes some shortcomings of the screen-out measure, particularly in the context of an operational system; the two measures may be used together for the most informative summary of screening performance. The measure has proved valuable in screen-set design for a connection table based graphical searching system and is suitable for long-term monitoring of system performance. It may also be used for comparison of screening procedures in different substructure search systems.

REFERENCES AND NOTES

(1) Lancaster, F. W. "Information Retrieval Systems: Characteristics, Testing and Evaluation", 2nd ed.; Wiley-Interscience: New York, 1979.
(2) Salton, G.; McGill, M. J. "Introduction to Modern Information Retrieval"; McGraw-Hill: New York, 1983; Chapter 5.
(3) van Rijsbergen, C. J. "Information Retrieval", 2nd ed.; Butterworths: London, 1979; Chapter 7.
(4) Sparck Jones, K., Ed. "Information Retrieval Experiment"; Butterworths: London, 1981.
(5) Cleverdon, C. W.; Mills, J.; Keen, E. M. "Factors Determining the Performance of Indexing Systems"; College of Aeronautics: Cranfield, England, 1966.
(6) Lynch, M. F.; Harrison, J. M.; Town, W. G.; Ash, J. E. "Computer Handling of Chemical Structure Information"; MacDonald-Elsevier: New York, 1971; Chapter 6.
(7) Adamson, G. W.; et al. "Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files". *J. Chem. Doc.* **1973**, *13*, 153-157.
(8) See, for instance, Adamson, G. W.; et al. "An Evaluation of a Substructure Search Screen System Based on Bond-Centred Fragments". *J. Chem. Doc.* **1974**, *14*, 44-48.

Combinatorial Problems in Computer-Assisted Structural Interpretation of Carbon-13 NMR Spectra

ALAN H. LIPKUS and MORTON E. MUNK*

Department of Chemistry, Arizona State University, Tempe, Arizona 85287

Received April 24, 1984

Combinatorial problems posed by a method for computer-assisted structural interpretation of ^{13}C NMR spectra based upon fragments consisting of a carbon atom and its α neighbors are discussed. The basic problem of generating all structures consistent with a set of inferred fragments that contains mutually exclusive alternatives is divided into two parts: generation of combinations of fragments and exhaustive assembly of each combination into molecules. Algorithmic solutions to both of these problems are presented in detail.

INTRODUCTION

The increasing importance of ^{13}C magnetic resonance spectroscopy as a tool in organic structure elucidation continues to stimulate the development of computer-based methods that aid the chemist in drawing structural inferences from the ^{13}C spectra of unknown compounds.[1] The most ambitious of these methods are those that automatically infer from the ^{13}C spectrum a set of possible substructures and then generate all candidates for the unknown that are consistent with these substructures as well as with the molecular formula and, if desired, information derived from other sources, spectroscopic or chemical. The basic strategy used is one of substructure inference followed by structure generation.

This strategy is part of a more general strategy employed by the chemist in the structure elucidation process. The computer modeling of this process is the goal of the evolving CASE system of programs.[2] The three major components of the system are spectrum interpretation (INTERPRET), molecule assembly (ASSEMBLE), and spectrum simulation and comparison (SIMULATE). The search for a convenient and efficient link in CASE between programs INTERPRET and ASSEMBLE has motivated the present work in combinatorial computing.

In general, computer systems for the structural interpretation of ^{13}C NMR spectra must use combinatorial algorithms capable of generating molecular structures from substructures that may overlap to some extent. Also, given that present-day knowledge of the relationships between molecular structure and spectral properties is incomplete and subject to inherent limitations, two or more substructures can arise from alternative interpretations of the same spectral feature. Thus, the algorithms used must be able to generate molecular structures from a set of substructures that may include mutually exclusive alternatives. Previous approaches have used different algorithms as well as different substructures.

In the CHEMICS system, designed by Sasaki et al.,[3] the selection of possible substructures based upon the ^{13}C spectrum of an unknown is accomplished by deleting from a list of small substructures (called components) all those that are shown by a correlation table of components and chemical shift ranges to be inconsistent with the spectrum. (The program can also use ^{1}H NMR, IR, and MS data to delete components.) From the set of remaining components, all subsets are formed that are consistent with the molecular formula and account for all ^{13}C resonances measured. From each subset of components, molecular structures are exhaustively generated by an appropriate algorithm,[4] thus producing all plausible candidates for the unknown compound. In later versions,[5] structure generation can be constrained by requiring the presence or absence of specific substructures.

Gray et al.[6] have developed a ^{13}C NMR interpretation method that uses a data base of substructures derived from a library of assigned spectra. Each substructure describes the environment of a particular carbon nucleus and is matched to the chemical shift and multiplicity of that nucleus. For each ^{13}C resonance in the spectrum of an unknown compound, those substructures that display the same multiplicity and a similar shift are selected from the data base. A reduction in the number of retrieved substructures is then attempted by an

0095-2338/85/1625-0038$01.50/0