	S/M
1A	0/6 - 3/6
1B	0/6 - 3/6
2	0/3 - 3/3
3A	0/12- 3/12
3B	0/12- 3/12
3C	0/12- 3/12
3D	0/12- 3/12
4	0/3 - 3/3

S/M = 0/24 TO 24/24

Figure 8. Range of scores by factor.

$$R_j = \sum_{i=1}^{4} W_i \bullet \frac{S_{ij}}{M_i}$$

SCORE:	0/24	1/24	2/24	3/24	...	8/24	...	12/24	...	15/24	...	21/24	... 24/24
RANK:	0.0	0.41	0.83	.124		.333		.50		.625		.875	1.0

Figure 9. Computation of ranks.

low or high scores and the majority of uses having various intermediate scores within a relatively narrow range. This result was acceptable and consistent with the goal of the scoring activity, which was to identify those chemical uses having the greatest potential for exposure. Whoever is responsible for selecting uses having the greatest exposure potential can establish any desired cutoff point along the curve. Should the entity conducting such a scoring and ranking operation desire to highlight one or more exposure parameters, weighting values can be assigned to selected parameters in accordance with the ranking formula. The result will be higher scores assigned to those uses characterized by the selected parameters of exposure. Furthermore, scoring for human and environmental exposure can be completely disassociated, if exposure to only one or the other is of interest.

Scores for the individual chemical uses can be made more accurate by devoting personnel time to accumulate more data in preparation for scoring. However, the basic concept of scoring and ranking chemical uses according to the methodology described above appears to be valid for its stated purpose.

REFERENCES AND NOTES

(1) U.S. Congress, Senate "Toxic Substances Control Act"; 94th Congress, 2nd Session: Washington, DC, 1976; S. 3149, Oct 11, 1976, Publication L. 94-469.

(2) Byer, W. L.; Landau, H. B.; Stalder, E. W. "Development of a Chemical Use Classification System to Facilitate Reporting under the Toxic Substances Control Act." *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 197-200.

Use of MACCS within ICI[†]

GEORGE W. ADAMSON, JOHN M. BIRD, GRAHAM PALMER, and WENDY A. WARR*

Pharmaceuticals Division, Imperial Chemical Industries PLC, Alderley Park, Macclesfield,
Cheshire, SK10 4TG England

Received August 19, 1984

ICI is developing a new system called SAPPHIRE—a user-friendly, interactive system for the storage and rapid retrieval of chemical structures and related property data, with interfaces to other systems such as molecular modeling, reaction design, and biological data handling. The chemical structure part of SAPPHIRE will be handled by MACCS software, written by Molecular Design Limited and enhanced by them in 1983 to meet ICI requirements for handling databases of over 400 000 compounds. Interfaces will use the Molecular Design program MACCSLIB.

INTRODUCTION

The Company Compound Center database shared by five ICI divisions contains about 360 000 compounds. Of these, nearly 190 000 are Pharmaceuticals Division compounds that have huge numbers of related biological test results. In addition, there is a file of commercially available compounds, and there are several smaller specialized files. Since the 1960s, all this data has been handled by the ICI CROSSBOW system,[1-6] with Wiswesser line notation (WLN) as the tool for structure representation. Chemists are unfamiliar with WLN and have therefore needed the intervention of information experts to search scientific data on their behalf. Moreover, structure display in CROSSBOW is a batch process, and the end-user has had to wait 24 h or more for his answers. His creative scientific ideas have therefore not been put to best use because his valuable train of thought is interrupted. In a truly interactive system this should not happen, and as each idea leads to another, the scientist should be able to pursue each train of thought to its logical conclusions when and how he himself wants.

ICI is therefore developing the SAPPHIRE system (*S*tructures *a*nd *P*roperties *P*roduced by *H*elpful *I*nteractive *R*apid *E*nquiry), a user-friendly, interactive system for the storage and rapid retrieval of chemical structures and related property data. The system will have interfaces to other ICI systems such as molecular modeling, reaction design, and biological data handling and even, it is hoped eventually, to systems external to ICI for handling information from the scientific literature.

The SAPPHIRE project is being developed by a team of about six analysts and programmers at ICI Pharmaceuticals Division but is funded by three of the five divisions who share the Company Compound Center database, namely, ICI Organics, Pharmaceuticals, and Plant Protection Divisions. The views of end users at all relevant ICI sites are being carefully considered in the system design.

The chemical structure part of SAPPHIRE will be handled by software written by Molecular Design Limited (MDL) and, in particular, by a version of MACCS enhanced to furnish rapid search on very large databases. This product is at present referred to as MACCS-BV (for "the big version" of MACCS) and is being beta-tested at ICI Pharmaceuticals Division. Later in 1984, MACCS-BV will be the standard version of

[†] Presented before the Division of Chemical Information, 187th National Meeting of the American Chemical Society, St. Louis, MO, April 11, 1984.

0095-2338/85/1625-0090$01.50/0 © 1985 American Chemical Society

MACCS available to all MDL customers, and the "BV" part of the title will be dropped.

ICI Americas has licensed a copy of MACCS-BV. The SAPPHIRE project also seeks to interface those compounds on the Company Compound Center database that have an ICI Americas reference number to a pharmacology system that is being developed at Wilmington, DE.

SOFTWARE

ICI is the first user of the MDL package MACCS-BV and was also instrumental in the appearance of the MDL product MACCSLIB. This is a library of routines that allows ICI to write FORTRAN programs that can access the MACCS database. Such programs will constitute the data-handling modules of SAPPHIRE. We are in the process of choosing a suitable database management system.

ICI has also licensed the MDL programs MARGEN and FORGEN, MOLRST and MOLSAV, and LAYOUT. The last three will be mentioned later in connection with structural data conversion. MARGEN and FORGEN allow the user to design output forms and plot structures and related data into suitable forms either on a terminal or as hardcopy. We are likely to use these programs only as an interim solution to data output from databases other than the large Company Compound Center one. In the long term, we shall almost certainly write our own output procedures. Our hardcopy requirements are very specialized, and our DEC LXY21 printer is not supported by MDL.

HARDWARE

A VAX 11/780 to be dedicated to the SAPPHIRE system was installed in Dec 1982. By 1985 or 1986, it is expected that there will be about 60 terminals on four sites in the UK and at ICI France. Access from Wilmington, DE, is still at a very early stage of development. Pharmaceuticals Division has a local area network, PLANET, and so far has two IMLAC terminals and 16 VT100 plus Retrographics terminals. The first eight terminals ordered for chemists all have local Anadex printers for screen dumps. There is one Versatec V80 printer-plotter driven by the VAX 11/780.

DATABASES

The following databases have been converted to MACCS: (a) the Cambridge X-ray crystallographic database and the related data; (b) the Hansch (Pomona college) physical chemistry database of structures and related logp data; (c) an internal ICI Hansch-type database of over 2000 compounds with measured logp values; (d) a database of compounds available commercially with supplier data and chemical names (CAOCI—the Commercially Available Organic Compounds Index—a misnomer since it also contains inorganics); (e) 310 000 compounds from the Company Compound Center database, with no related data at present other than a numerical key.

SAPPHIRE is being developed in phases, and at present, the Company Compound Center database under MACCS on the VAX machine is purely a search database. Registration is still done, with WLN, on-line to a Burroughs mainframe. Every evening, the day's changes to the Burroughs database are transferred to the VAX by magnetic tape and converted to MACCS on the VAX. Eventually, it is intended that both registration and search will take place on the VAX.

The initial exercise of converting 301 537 structures proceeded remarkably smoothly in 83 runs of 12–14 h on the VAX in an elapsed time of just under 4 months. Over 290 000 of these structures were scanned by information scientists, 5600 were hand-drawn from scratch, and about 11 000 had their structure modified. These figures should be regarded with caution since databases vary and opinions vary on what constitutes a pleasing structure.

The reasons why conversion progressed so efficiently could be summarized as follows. ICI has always taken pride in the accuracy and consistency of its WLN encoding, has adopted useful conventions, such as the CROSSBOW "double-ampersand" suffix, and has avoided the pitfalls of contractions and multipliers. All WLNs have been subjected to stringent manual and machine tests. This has, without a doubt, minimized crashes of the conversion software and reduced the number of structures needing correction. Four months were spent planning the whole conversion exercise and writing software to provide the highest possible degree of mechanization. Thus, selected data was transferred from Burroughs to VAX and VAX to Burroughs every night. Conversion was run on the VAX. Converted structures were divided into three categories: no structures, bad structures found programmatically, and all others (which needed scanning). Files of structures for scanning were automatically set up. The reference numbers for bad structures found during scanning were held in separate files. Each day CROSSBOW 9 in. × 5 in. structure display cards were output for no structures, bad structures found programmatically, and bad structures found by scanning. Two preuniversity students copied structures from these cards onto IMLAC screens. Six or more information scientists were organized in shifts to scan 290 000 structures—at six structures to a screen in MACCS PLOT mode, one person could scan 500 structures in 0.5 h. In addition to the scanning exercise, all structures drawn from scratch were checked by an information scientist. Two IMLAC terminals were in constant use from 8:30 a.m. to 5:30 p.m. every day.

Numerous CROSSBOW and utility programs were used to perform all the automatic data manipulation on the Burroughs machine. The main VAX programs were CTGEN (written by ICI) and LAYOUT and MOLRST (supplied by MDL). CTGEN calls on the DARING routines (marketed by Fraser-Williams Scientific Systems) to convert a WLN to a connection table. CTGEN also makes the connection table suitable for input to LAYOUT, which generates structure coordinates. MOLRST updates a MACCS database.

When LAYOUT was written in fall 1981, it was probably MDL's first fast response to the specialized needs of ICI. Some 450 000 structures were run through LAYOUT at ICI Pharmaceuticals Division in 1983. The remarkably efficient combination of CTGEN, DARING, and LAYOUT has been one of the success stories of the year.

MACCS-BV

In the early days of its development, MACCS did not meet ICI's requirements for a truly interactive system in that search times would sometimes be too long on a database of 400 000 compounds. When ICI licensed MACCS in 1982, it was part of the agreement that MDL should supply "fast MACCS", or MACCS-BV, capable of working on databases of over 400 000 compounds, before mid-Sept 1983. MACCS-BV was in fact installed in July 1983, 2 months ahead of schedule. It was also 2–3 times faster than the requirements specified by ICI. It is difficult to express this in terms of precise search times since databases vary, searches vary, and the configuration and usage of the machine affect the elapsed time. Suffice it to say that MACCS-BV is in the same league as CAS ONLINE or DARC as regards search speeds.

MACCS-BV is compatible with standard MACCS in that MACCS-BV can be used for search and registration on either BV or standard databases. It also has the full data-handling capabilities of standard MACCS. Both these facilities were unexpected bonuses, in that ICI had not specified a requirement for them.

Unfortunately, we were not able to implement the new package until Jan 1984, largely due to factors beyond the control of Molecular Design. During 1983, a second generation of VT100/Retrographics terminals was released and changes in MACCS were necessary. In the summer and early autumn of 1983, our priorities did not involve MACCS-BV and we told MDL we would not need the new version of MACCS-BV until November. Looking back on it, we now wish we had asked for an earlier release. Further delays were caused by a supplier being unable to obtain the required terminals and by a new release of the local area network firmware for Pharmaceuticals Division.

Despite the delays, as of April 1984, we *have* run five basic training courses and one advanced one on various ICI sites. About 40 users have been trained, and MACCS-BV is now being accessed from 24 or more terminals. We think that our machine, as presently configured, will handle about 10 simultaneous MACCS users, but we are starting to reevaluate our hardward requirements for the future. The number of users is increasing rapidly, and security, passwords, and access control are being given urgent consideration.

CONCLUSIONS

In the early days of SAPPHIRE, we were faced with the choice of writing a system ourselves (as we did with CROSSBOW) or buying a suitable software package. Although the expertise was available in-house to write a system, it was quicker and cheaper to buy an externally written package. It is an indisputable fact that we could not have achieved all the 14 months' progress reported here without the use of MDL software. However, since MACCS was not designed specifically to meet ICI's in-house requirements, it is not surprising that there have been suggestions for additional improvements.

Information scientists have suggested enhancements, especially those that would make MACCS query formulation more flexible, so that fewer queries required multiple searches. (The latter improvement they are likely to get in 1985.) End users are more concerned with making MACCS even more user friendly.

The concerns of systems personnel are bugs, operability of the package as a unit in a multiuser environment, and running batch jobs, essential for large files. However, ICI is a testing site for linking and using various MDL programs within a multiuser environment with a very large database. Some problems were therefore expected, and MDL has been very responsive to all our demands. They have set up a Quality Control Division, and in the few cases where a bug (or even a feature) has caused us serious operability problems, Molecular Design has moved swiftly and replacement software has, if necessary, been sent within days by air courier. The considerable geographical distance between ourselves and MDL and the time difference have never been a major problem. All questions and inquiries are answered with remarkable promptness by the customer service personnel at MDL, and we are sure that there have been cases when programmers have burnt the midnight oil in order to accommodate us.

At this stage we are not able to reveal fuller details of SAPPHIRE design. It is not the object of this paper to discuss database management systems or how the in-house software we are developing will be tailored to user needs. MACCS is a system of considerable interest to the information scientist, but there have been very few publications on it.[7-9] We are pleased to submit an early one, and we shall publish further details at an appropriate time.

REFERENCES AND NOTES

(1) Hyde, E.; Matthews, F. W.; Thomson, L. H.; Wiswesser, W. J. "Conversion of Wiswesser Notation to a Connectivity Matrix for Organic Compounds". *J. Chem. Doc.* **1967**, *7*, 200-204.

(2) Thomson, L. H.; Hyde, E.; Matthews, F. W. "Organic Search and Display Using a Connectivity Matrix Derived from the Wiswesser Notation". *J. Chem. Doc.* **1967**, *7*, 204-207.

(3) Hyde, E.; Thomson, L. H. "Structure Display". *J. Chem. Doc.* **1968**, *8*, 138-146.

(4) Eakin, D. R. "The ICI CROSSBOW System". Ash, J. E. "Connection Tables and their Role in a System". In "Chemical Information Systems"; Ash, J. E.; Hyde, E.; Eds.; Horwood: Chichester, England, 1975.

(5) Eakin, D. R.; Hyde, E.; Palmer, G. "The Use of Computers with Chemical Structural Information: ICI CROSSBOW System". *Pestic. Sci.* **1974**, 319-326.

(6) Townsley, E. E.; Warr, W. A. "Chemical and Biological Data—An Integrated Online Approach". *ACS Symp. Ser.* **1978**, *No. 84*.

(7) Wipke, W. T.; Dill, J. D.; Peacock, S.; Hounshell, D. "Search and Retrieval Using an Automated Molecular Access System". Paper presented at the 182nd National Meeting of the American Chemical Society, New York, Aug 1981.

(8) Wipke, W. T. "MACCS and REACCS". *Proc. Soc. Polym. Sci. Jpn.* **1983**, 14-19.

(9) Warr, W. A. "MACCS—An ICI View". Proceedings of the International Online Information Meeting, 7th, London, Dec 6-8, 1983.

Monte Carlo Studies of the Classifications Made by Nonparametric Linear Discriminant Functions. 2. Effects of Nonideal Data

TERRY R. STOUCH and PETER C. JURS*

The Pennsylvania State University, University Park, Pennsylvania 16802

Received December 20, 1984

Recently, the levels of correct classifications due to chance that were attainable by nonparametric linear discriminant functions (NLDFs) were studied. That previous work dealt with easily generated, idealized data. Because of this, the application of those results to actual studies using nonideal data may not be warranted. The studies reported here analyze the effects of zero values, indicator values, and multicollinearities: variations that occur in actual data and that could affect the levels of random classifications. Three structure-activity relationship studies that were performed with NLDFs are also examined.

Discriminant functions can be visualized as surfaces that divide a data space into different regions. The aim of this method of pattern recognition (PR) is to divide the data space into regions of significance. For example, in a structure-activity relationship (SAR) study the data space would be populated by points, often referred to as patterns, which represent compounds with interesting biological activity. A useful discriminant would divide the data space into regions

0095-2338/85/1625-0092$01.50/0