Installation and operational experiences with MACCS (Molecular Access System)

D.J. Polton

Historical background

Computerised chemical substructure searching began at Shell Research Sittingbourne during the 1960s with a file of chemical structures coded in string form using the IUPAC (Dyson) chemical notation. Work on the coding of structures in this notation had begun about 1961 when the definitive rules were first published, and whilst the file was being created it was also being searched for compounds containing similar chemical fragments by using once-off programs. In order to evolve into a fully fledged computer system, certain drawbacks in the notation had to be overcome and, based on a modified notation [1], a series of programs had been written by 1966 to enable registration, substructure searching, molecular formula checking and other activities to be carried out [2]. After many years of use, it was decided that a move towards an interactive graphics system should be made, and several ideas were followed up. When, during these deliberations, Molecular Design Limited (MDL) announced their product MOLEX this was examined and assessed to be a suitable replacement for the old system provided that the structures and data already stored could be satisfactorily transferred. Molex, which is now known as MACCS (Molecular ACCess System), was subsequently purchased and has now been in use at Sittingbourne Research Centre for a year and a half.

With the arrival of MACCS at Sittingbourne Research Centre in May 1980 a new era was about to begin. The ever-widening gap between the closure of registrations into the previously used system and the installation of its replacement, due mainly to machine failures and staffing problems, was becoming quite a serious matter. New compounds had not been registered for several months by the time MACCS became available, and this had created a substantial backlog. This, in turn, made it possible to carry out thorough checking of compounds received from external sources for presence in the file. It was therefore imperative to install the new system and make it operational as soon as possible.

The author is at Shell Research Ltd., Sittingbourne, Kent. England. *This paper is based on one given at the 5th International Online Information Meeting, London, December 1981.*

About the author

Donald J. Polton

Donald Polton is a graduate in chemistry from the University of London. He spent several years as an organic synthesis chemist before joining Shell Research Ltd. in 1963 in the field of computing and information processing. In this position, he was responsible for the development of many early computer systems, including that based on Dyson (IUPAC) chemical notation.

Since 1979 he has been secretary of the Chemical Information Group of the Royal Society of Chemistry. Private interests include horticulture, photography and philately.

Installation

When MACCS was installed there were no major problems. The work was done quietly on a Bank Holiday Monday, and the system was available for demonstration to staff on the following day, May 6th, with a test file of 12,000 structures provided by the suppliers. In fact, loading MACCS onto a Prime computer has, in my experience, never been a problem. It was done at Leeds University in September 1981 for the Autumn meeting of the Royal Society of Chemistry, and the system was operating within a few minutes. It is also possible to load MACCS on a VAX 11/780. In this case I cannot speak from experience and can therefore make no comment. With the Prime computer it is merely a case of having an area on a Master File Directory (MFD) and sufficient space for the data for MACCS to become usable immediately. The programs supplied go into a user file directory (UFD) known as PROGS from which they are operated by using the statement RUN. Associated HELP files are also loaded, together with the periodic table and other reference files. A file of system line numbers, indicating spooler type, and which have graphics terminals and/or Tektronix 4662 plotters attached has to be set up, and the directory of users with authority levels attached to their log-in names must be prepared. Certain users at Sittingbourne are permitted to write structures and data to the file, but few have authority to alter or delete. This security aspect of MACCS is very easily set up and the user file created may be edited by those with authority to do so. No person can access a protected MACCS database without prior entry into the user file. For certain purposes it may be necessary to bypass the molecular formula check during registration, such as when entering large numbers of structures from a command file. A temporary amendment to the user file will enable this to be done. Apart from certain data types which are to

appear on the screen with the structure, such as the registry number and the date, data files may be added at any time after installation. Even after installation of these, it is possible, as was done at Sittingbourne, to carry out a file reformat process.

File conversion

After the installation of MACCS at Sittingbourne, the file of proprietary Shell compounds was to follow in batches. This was being converted into MACCS format from the file of structures produced using a chemical structure typewriter, i.e., the keystroke images. The alternatives available, the chemical notation previously mentioned and connection table records, were not used. Unfortunately, the conversion took longer than had been expected. The major reason for this was that the suppliers of the conversion program did not have full information on the method of construction of the records, in which a great many abbreviated forms of structural fragments were used. Even when the structures had been fully processed there was a proportion of records which had failed to convert, and therefore had no structure. More seriously, a large number had an incorrect structure. These latter consisted principally of compounds with methyl groups missing, in which the methyl group had been originally typed as Me rather than CH3, an abbreviation which had not been allowed for in the original conversion. The complications caused by structures being entered incorrectly are cumulative. If a structure with a methyl substituent is registered as unsubstituted, then when the genuine unsubstituted material is registered it is regarded as already in the file and is rejected. Strange to say, the discovery was not made for some months. In a very short time much followed in running programs to re-read the original typewriter stroke file, and checking the results. Even after this a number of structures had to be entered manually, as the original tape contained many records which were unreadable. Eventually, a reliable working file was created, maybe with fewer errors than would otherwise have been present, for inspection of the data had been very close following the discovery of incorrect structures, and some other anomalies had been discovered as a result. Furthermore, to detect as many errors as possible, a check was carried out between the molecular formulae on the original file and that calculated by MACCS from the compound structure. This proved very enlightening in that there were many hydrogen count errors in the original formulae, which had been worked out manually from the structure in days gone by.

One peculiarity of the original notation file had always been the use of more than one reference number to register the same compound. The reasons for this are largely historical, compounds having been obtained from various laboratories and renumbered on arrival into the site series for the purpose of biological testing, but often referred to under the original number. Also, with incomplete data card files, a chemist about to synthesise a compound could not know that it had already been tested under another number. This was before the days of computerised files, of course. Included in the development of the notation based system was a complete cross index file which would prove very useful to explain in the MACCS file the reason for so many blank screens where structures should have been expected, for a structure could at that time only be entered once into MACCS, and subsequent

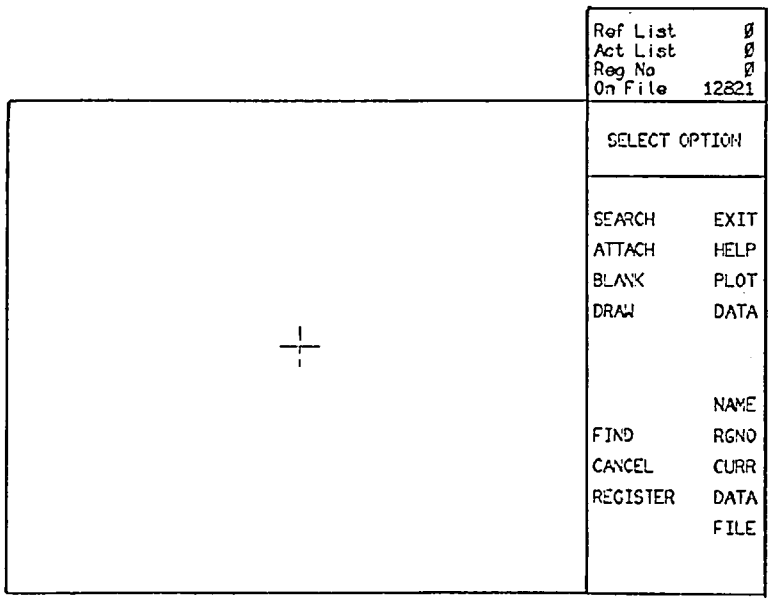

Fig. 1. *Executive mode menu*

entries could only be made (for the purpose of registering later batches of the material) by registering a 'blank screen'. It was a relatively easy matter to convert this file into MACCS data format by the use of the Prime data editor and transfer it directly into the new database.

Testing

So much for the installation. With the arrival of MACCS a programme of thorough testing was devised and put into effect before the system could be called 'operational'. Every aspect was carefully examined to ensure perfect operation, this being especially important with a new product of which Shell Research were the first European buyer, and the third buyer worldwide. In the event this proved to have been a wise undertaking, for a number of problems were encountered which were passed to the suppliers and dealt with by them. Suggestions for the improvement of MACCS were put forward, some of which were incorporated in a subsequent revision. Now, one and a half years after installation, we have a system used by over eighty members of the staff, which is easy to manipulate and which is, we hope, trouble free. Some more minor improvements and modifications are still awaited and the arrival of MACCS version 3.0 is eagerly looked forward to in 1982.

The system

A few words about MACCS itself must be made, although this is not intended as a descriptive paper. The Executive Screen of MACCS, seen when first logged in, is the key to all other processes (Fig. 1). It is also the major area for carrying out registration processes. Some of the options lead to other modes — DRAW, ATTACH, SEARCH and DATA, and others are for file handling. The words on the left

```
BUTYRYLCHOLINE IODIDE, 99%
DPMACCS 02102881113620 1   0.00601      0.00000     71
ALDRICH
 13 11  0  0  0
    2.5972    0.8537    0.0000 C   0  0  0  0  0
    4.1483    0.1563    0.0000 N   0  3  0  0  0
    1.3287   -0.1082    0.0000 C   0  0  0  0  0
    4.6954    1.5992    0.0000 C   0  0  0  0  0
    5.6693   -0.2766    0.0000 C   0  0  0  0  0
    3.6072   -1.2565    0.0000 C   0  0  0  0  0
   -0.0721    0.7455    0.0000 O   0  0  0  0  0
   -1.4249   -0.3607    0.0000 C   0  0  0  0  0
   -2.8377    0.2285    0.0000 C   0  0  0  0  0
   -4.1663   -2.0621    0.0000 O   0  0  0  0  0
   -3.9679   -0.8116    0.0000 C   0  0  0  0  0
   -5.2605   -0.0481    0.0000 C   0  0  0  0  0
   -7.3106    1.3407    0.0000 I   0  5  0  0  0
  1  2  1  0  0  0
  1  3  1  0  0  0
  2  4  1  0  0  0
  2  5  1  0  0  0
  2  6  1  0  0  0
  3  7  1  0  0  0
  7  8  1  0  0  0
  8  9  1  0  0  0
  8 10  2  0  0  0
  9 11  1  0  0  0
 11 12  1  0  0  0

OK,
```

Fig. 2. *Standard molecular utility (SMUT) file of butyrlcholine iodide*

and right are used in pairs, for example to register a displayed structure REGISTER and CURRENT are used, and to retrieve a structure FIND, in combination with CURRENT if the structure is displayed, REGNO if its registry number is known, or FILE if it exists in a Standard Molecular UTility (SMUT) file external to MACCS (Fig. 2). Any structure on display may be registered into a SMUT file by using REGISTER FILE, these being useful for conversion into 3D images with other programs.

In normal usage of the graphics terminal, options are selected in executive mode with a light pen. If no structures are to be drawn, it may be preferred to work from the keyboard, keying in initial letters of the various options. Interchange between graphics and keyboard modes is easily effected by keying the letters G and K respectively. It is possible, given a parent compound, to draw and register compounds from the keyboard using ATTACH mode, a facility which builds up structures from atoms and stored templates, and which is becoming more popular as staff discover its ease of operation. To become more amenable, certain improvements to ATTACH mode such as group deletion have been requested, and it is hoped that these will become available soon. But this will never replace DRAW mode, with its many options, including GROUPMODE, for manipulating the whole or part of a structure whilst on screen, and MOVE, enabling any atom to be moved to any other position on the screen. Indiscriminate use of the CLEAN option in DRAW mode, which tidies up a badly drawn structure, does mean, however, that structures must be examined and all too often altered by a central agent.

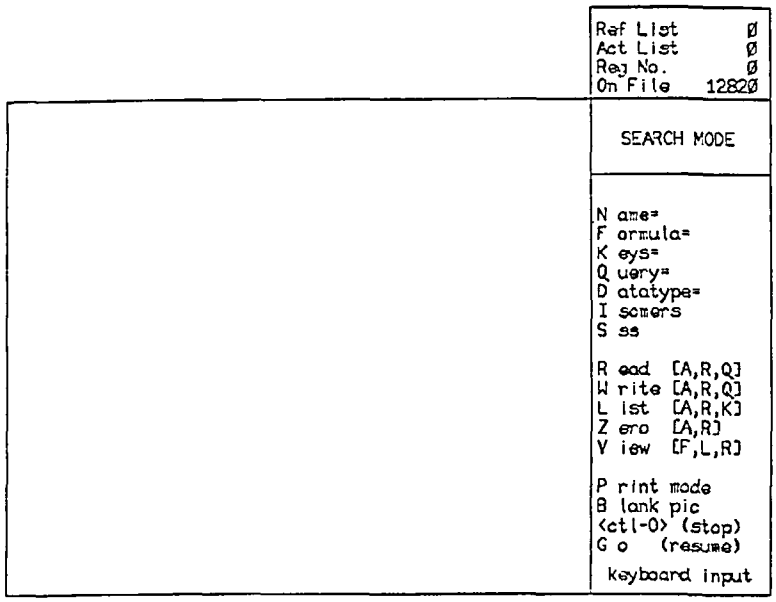

Fig. 3. *SEARCH mode screen*

This should be used repeatedly during the process of drawing, and not after the molecule has been drawn, as for the more complex structure the result may be unacceptable.

Of the many SEARCH facilities available (Fig. 3), substructure searching is the most important to the general user. Name, formula and data searches are sometimes done, but more by the MACCS administrative and computing staff than elsewhere. Substructure searches are not difficult to operate. A general formula may be drawn using SSS (SubStructure Search), an option in DRAW mode which allows AND, OR and NOT logic to be applied to atoms and bonds. Great care must be taken in establishing the current type of bond in cases where ring closure is permitted. A search on the diphenyl molecule will not find phenanthrene unless the connecting bond is given the option of being aromatic.

A MACCS substructure search is a two stage process. The KEYS are previously recorded fragments, the presence of which is recorded against each compound in the file during registration. The primary search for keys filters out all of those structures not containing the various key groups as defined by the search question, and reduces the file considerably before the substructure search proper on the connection table begins. When setting up a file, the user may choose to use his own search keys rather than those supplied. When devising keys, caution must be exercised to avoid missing structures during a search. It must be remembered that a key is switched on if it fits the search question, and if the question is general and the key specific then those structures not fitting the specific key, although covered by the general question, will be lost. For example, key is CH3, question is C, the structure with CH2 but not CH3 will not be found. Searching is very straightforward

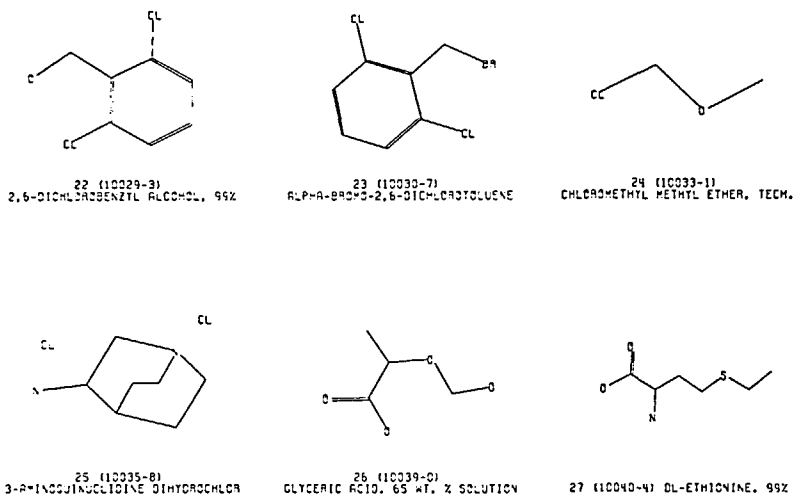

Fig. 4. *VERSATEC plots*

and reliable provided that care is taken in drawing the general structure or search question. It is possible to check results while the search is in operation, to carry out a secondary search on the result file, to see how the result file is growing and take appropriate action, returning to refine the question if too much is being found. Manipulation of different search result files, such as intersection, addition and subtraction may be performed. In this way a search may be carried out for compounds containing one substructure but not another. Hard copy of structures either by registry number, or as the result of a search, may be obtained on Versatec or Tektronix 4662 plotters at up to six per page (Fig. 4). With the latter, additional text may also be keyed in.

In DATA mode as many different data types as required may be stored and searched (Fig. 5). Data files may be taken from the system, edited using the Prime data editor, and then returned. Great improvements to the range of data the system can hold are expected in the near future, but the present version of MACCS permits textual and numeric data, formatted and unformatted, to be held, with the facility of doing range searches between two quoted numbers. The facility to hold arrays is promised in the next update. Data searching is usually associated with requests for information on specific compounds or those obtained by substructure search rather than the whole file, and is in this case quite rapid.

Last but not least is the link to the 3-dimensional display programs DISP, SPACFIL and ORTEP via the energy minimisation program PRXBLD, which works very well for certain types of compound. These are supplied as an addition to MACCS and are accessed via the SMUT file: It is expected that the future will bring forth great improvements in 3-dimensional structure representation. For the present, MACCS provides a very valuable interface to this type of program. Not only this, it can be linked to programs which calculate physical properties such as partition coefficient and vapour pressure.

```
32044 ENTRIES IN DATA BASE
MACCS: Data
DMODE: Search data type(s) = 20
Search string: 1.43
 > 298    (10491-4 ) <REFRACTIVE.INDEX>                          DT0020
          1.43

 > 766    (11115-5 ) <REFRACTIVE.INDEX>                          DT0020
          1.43

 > 819    (11180-5 ) <REFRACTIVE.INDEX>                          DT0020
          1.43

 > 3451   (14724-9 ) <REFRACTIVE.INDEX>                          DT0020
          1.43

 > 7074   (19477-8 ) <REFRACTIVE.INDEX>                          DT0020
          1.43

!<INTERRUPT>
DMODE: Exit
#MACCS:Exit

OK
```

Fig. 5. *Data search for compounds of refractive index 1.43*

To sum up, MACCS is very easily installed on the Prime series of computers. Once set up, with a good database, its operation is very quickly learned, even by those with little or no knowledge of chemical structure representation. It has proved to be a very satisfactory system. There is of course scope for improvement, but much of this may be at the expense of efficiency. Perhaps future improvements in computing will make feasible some of the options which at present are wishful thinking.

References

[1.] D.J.Polton: *Inf. Stor. Retr.*, 1969, **5**, pp. 7–25.
[2.] D.J.Polton: *Inf. Stor. Retr.*, 1972, **8**, pp. 191–201.