COMPUTER ANALYSIS OF CHEMICAL REACTION INFORMATION FOR STORAGE AND RETRIEVAL

A study submitted in fulfillment of the requirements for the degree of Doctor of Philosophy of Information Science at the University of Sheffield by Peter Willett, BA, MSc.

October 1978

Summary

This thesis addresses itself to the automatic generation of machine readable descriptions of the substructural changes occurring in chemical reactions; the primary aim of the work is the inclusion of such data in current computer based chemical information systems.

In the first chapter a review is given of the methods, both manual and automatic, which have been described for the indexing of chemical reactions. In particular, a critical evaluation is made of the work on automatic reaction indexing carried out in Sheffield over the past decade. This work, which used both Wiswesser Line Notation(WLN) and connection tables as the structure representation, has formed the basis for the multi-level, whole structure, WLN fragmentation procedure described in the second chapter. The basic fragmentation algorithms are outlined together with their implementation in a program for producing printed reaction indexes. Experience of the use and retrieval effectiveness of such an index is compared with that of a commercially available reactions documentation service.

Whilst of wide applicability, the use of a notation leads to difficulties both in computer processing and in the quality and level of description of the analyses produced. Many of these problems are eliminated if connection tables are used as the structure representation and in the third chapter we present an approximate structure matching algorithm which enables the rapid identification of many of the subgraph isomorphisms present between the sets of reactant and product molecules in a reaction. The technique is based upon an adaption of the Morgan algorithm to the description of circular substructures and this has permitted the development of a systematic method for the selection of

fragments as screens for use in chemical susbtructure search systems.

Finally an experimental reactions retrieval system is described which uses both the methods of analysis described earlier to characterise the reactions in the search file. A range of reaction queries have been put to the system with reasonable results in terms of the material retrieved. The techniques could be easily implemented in a conventional substructure search system.

(214 references)

Ackowledgements

My grateful thanks are due to the following people and organisations:

(i) Prof. Michael F. Lynch for his enthusiastic support of my research

(ii) the Department of Education and Science for the award of an Information Science Research Studentship

(iii) the Institute for Scientific Information for the tapes from which the reactions data base was created

(iv) the Data Services Section, Imperial Chemical Industies (Pharmaceuticals Division) for the provision of software

(v) the members of the Research Information Department, Pfizer (UK) Limited in general, and Dr. David Bawden in particular, for their collaboration in the WLN - Derwent comparison, for a set of user reaction queries and for many helpful discussions

(vi) the members of the Computing Services Department, University of Sheffield for their help in the development and running of my computer programs.

Statement of responsibility

With the exception of standard ICL software, all of the programs designed and implemented in the course of this study have been the work of the author alone. The WLN - Derwent comparison was a collaborative project and, accordingly, has not been included in the main body of the thesis.

Contents

CHAPTER I

Previous work in the field.

I.1 Introduction

The advent of the computer has led to rapid changes in the methods used for the analysis, storage and retrieval of (primarily) scientific and technical information(1). The primary application has been in the rationalisation and mechanisation of the procedures necessary for the publication of printed secondary sources of information i.e. indexing and abstracting tools(2). Secondly, the availability of a machine-readable form of the source material, the data base, has allowed the development of computer based information services which provide users with a variety of access mechanisms by which they may interrogate the data base. Batch processed current awareness and retrospective search facilities became generally available about twelve years ago in the shape of tapes from Chemical Abstracts Service(CAS) and the National Library of Medicine(3,4). The rapid development of telecommunication networks and of disc storage and multiprogramming technology has meant that both of these functions are now frequently performed using online systems which allow both a greater immediacy of response and the potential for more refined search techniques(5); online searching has indeed proved so popular that the continued production of the source hard copy publications has been called into doubt(6).

It must be emphasised that these systems involve operations primarily upon the form of records i.e. manipulations with (sub)strings of alpha-numeric characters in specified data elements of the document file. The decisions as to which character strings should be considered as representing the content of the document are, in large part, still performed manually although research in the field of 'content analysis' i.e., the automatic indexing of natural language documents, is being carried out by many workers(7). In the case of chemical structure information, the differentiation between form and content is much less well defined since the form, the structure diagram, is a much closer

representation of the content, the wave equations describing the molecule, than in the case of natural language words. It is primarily for this reason that computer based information systems are, perhaps, most widely established in the field of chemistry(8,9,10,11) where although many of the items to be handled are textual or numerical in nature, the heart of a system is the chemical structure file which contains the machine readable representations of a large number of chemical compounds. The development of methods for the notation of molecules has been a long, not to say tortuous, process(12,13,14) but in the context of this dissertation we shall be concerned mainly with three of the methods that have been used to describe compounds in a machine readable form: these are fragment codes(15), connection tables(16) and Wiswesser Line Notations(WLN)(17,18). The two largest compound files are those belonging to CAS which contains over four million compounds in connection table form(19,20) and to the Institute for Scientific Information which contains about three million WLNs(21). Commercial files, such as are operated by the research departments of many agrochemical, pharmaceutical and petrochemical firms, are substantially smaller but may well contain over fifty thousand compounds(22,23,24,25). These files, together with the systems that control the storage and retrieval of information from them, represent a considerable investment of time, money and expertise and it would seem worthwhile to consider other uses to which they might be put.

This thesis considers the application of current structure handling techniques to the provision of rapid and easy access to chemical reaction data(26,27). This is of fundamental importance to several branches of chemistry(28) but we shall be primarily concerned with the field of synthetic organic chemistry where the need for adequate means of retrieval has been apparent for many years; the preface of the first edition of Weyl's famous book on organic chemistry contained

the statement that a scientist could hardly hope to be familiar with every one of the innumerable methods described therein(29). More recently, both Meyer(30) and Valls(27) have called attention to the importance of providing adequate reaction information; it has been stated that approximately one half of all the organic queries put by chemists to the BASF Ludwigshafen Documentation Group were concerned with reactions(31). As there are now over four million compounds known and any one may be transformed into many others by suitable reactions, it can be seen that the amount of potential data is enormous and it is also constantly increasing(32,33). Hendrickson, indeed, has pointed out that there are large classes of reactions for which there are as yet no known members(34). There are often many ways in which a molecule may be synthesised and yet there are currently few aids to help the chemist in his search for a viable synthetic pathway. The difficulty of the problems involved may be evidenced both by the wide recognition of the achievements of chemists such as Corey and Woodward and by the frequent use of terms such as 'elegant' in reviews of syntheses: synthetic organic chemistry has indeed been described as "an art in the midst of a science"(35).

It might have been expected that computers, with their ability to compare and collate large volumes of data would provide a ready means for the control of chemical reaction data but this has not proved to be so. At least in part, this lack of success has been due to the limited amount of research carried out in the field - the documentation of a reaction presupposes a method for the encoding of the reacting molecules, or some portion of them, which has only become feasible within the last ten years or so - but the main problem, as has been pointed out again and again(26,27,36,37,38,39,40), is that whereas a chemical molecule is a unique entity and thus susceptible to listing in a canonical form, such as <u>via</u> the CAS Registry System(20), a reaction

has many parts, all of which may need to be stored for subsequent retrieval. The large number of characteristics - starting materials, products, reaction sites, catalysts, conditions, bond changes and yields - makes the organisation of the information and the selection of suitable data items quite difficult(37).

There seems to be fairly general agreement that at least the following four data elements should be present in a reaction file if it is to be capable of handling a reasonable range of query types(27):

(i) compound information: ideally this should include details of any intermediates formed in the course of the reaction but in general the reactants and products alone will be encoded.

(ii) experimental conditions: these include such things as catalyst, concentration, temperature and solvent.

(iii) reaction analysis: a definition of the changes occurring during the course of the reaction.

(iv) bibliographical details.

Of these the reaction analysis is both the most important and the least well defined. Outside of this department, almost all of the systems that have been examined or implemented to data, however effective in operation, have been very expensive to create and use since the reaction analysis has been performed manually. A large scale reaction file can only be efficient in operation if the analysis for storage and retrieval can be performed automatically. There are thus valid economic grounds for the development both of general techniques of representation and of software systems for "in house" processing; were such packages to become generally available, it seems probable that machine readable reaction data would become available from commercial sources such as ISI. The availability of large machine readable reaction files would also form a natural complement to the rapidly growing area of computer aided synthesis design(41). Since the

potential reaction data base is potentially very large, reaction indexing programs must be simple in concept and efficient in operation if economical processing rates are to be achieved; synthesis programs, on the other hand, perform sophisticated manipulations using a limited file of, perhaps, three hundred basic reactions(42,43,44). A potentially useful approach would be to use the transformations output by a synthesis program as the basis for searches in a more substantial reactions file.

I.2 Manual methods of indexing reactions

In the next two sections we shall consider some of the many methods that have been proposed for indexing chemical reactions; for the present we shall limit ourselves to those where the intellectual tasks of analysis and representation have been performed manually, albeit for subsequent mechanised storage and retrieval in some cases.

As with compound information, the earliest forms of reaction indexing were based upon nomenclature and to this day the most widely employed and most easily understood description is the use of a trivial name, usually that of the chemist(s) who originally discovered the reaction. Terms such as Skraup synthesis, Claisen condensation and Clemmensen reduction are common in the literature and several compendia are available, the most comprehensive of these containing several hundreds of entries(45,46,47,48). Nomenclature may occasionally prove very powerful in rapidly describing complexes which can be difficult to characterise using more systematic methods e.g. the Cope rearrangement. Generally, however, the use of indexing terms which have no direct relationship with the reaction that they are supposed to describe may lead to severe problems in retrieval. Thus structurally similar transformations may be separated which might be considered more fruitfully in conjunction and, as was pointed out by Clews(49), there may also be disagreement as to the exact extent of the reactions that should be considered under a single heading. However the greatest deficiency is simply the lack of coverage offered by such a system since the overwhelming number of reactions have not been graced by a suitable appellation.

A more systematic use of nomenclature has been suggested by Patterson and Bunnett(50) and by Kunz(51). The former authors proposed that the name of a substitution reaction should be composed of the name of the incoming group, the syllable 'de', the name of the

outgoing group and the suffix 'ation': thus the hydrolysis of an alkyl chloride would be called hydroxydechlorination. Vleduts has shown that the scheme is ambiguous even for some cases in the limited field of functional group interconversion reactions(52) and the authors give no indication as to how the system could be extended to cover more complex processes. Nomenclature has also been used by Ursprung-Fischer who found that the uneven distribution of reaction types amongst the classes of a proposed notation scheme for reactions necessitated the subdivision of several of the classes by the use of trivial nomenclature(53). Dyson and Riley described reactions by a mnemonic code descriptive of the reaction type e.g. Chl-01 represented chlorination(54). None of these schemes appear to have been used in practice although the International Union of Pure and Applied Chemistry (IUPAC) has recently shown interest in an extended version of Patterson and Bunnett's system.

Of greater practical importance is the annual publication "Organic Syntheses" which is devoted to the description of the preparations of specific compounds so that the indexing is primarily upon the basis of the name of the product(55). Mischenko has described an index to the Russian translation of this publication in which broad classes, such as halogenation or nitration, are subdivided by a structural expression of the particular reaction class(56). Structure based searches are the prime means of access to Chemical Abstracts and Current Abstracts of Chemistry(CAC) for reaction queries; in the latter case an automatic search is also possible since a WLN magnetic tape is available which contains a list of all the new compounds(21) whilst limited substructure searches of CAS files may be performed online using the CHEMLINE file(57). It should be noted that there is no inherent difference between a systematic name and a unique linear notation so that compound names may become widely used as a machine readable structure representation.

Work is currently being carried out, primarily by CAS, upon the use of systematic nomenclature for structure storage and search(58,59) and algorithms are being developed to generate connection tables from an input compound name(60,61). Applications of this research to the area of reaction indexing are not impossible(62).

It is convenient at this point to mention the use of indexes of functional groups and of reagents. The former are usually arranged by the functional group of the product and then subdivided by the functional group of the reactant which has been involved in the change. Obviously, such an approach can only deal satisfactorily with simple changes, especially if the reacting molecules are polyfunctional(29,63, 64). Examples of reagent based indexes are "Synthetica Merck" and the well known Fieser and Fieser(65,66); under each reagent is listed the types of reactions for which it may be employed, usually with details of the appropriate reaction conditions.

A more systematic approach is to classify reactions according to the bonds broken or formed in the course of the reaction, an idea first proposed by Weygand(67). Theilheimer developed Weygand's system to produce a simple classification based on the types of bonds broken and formed and on the nature of the reaction(68). Reactions are described by a three part symbol string; the first part refers to the bond formed in the reaction, the second is a bond change indicator and the third the bond broken. The indicators represent addition, rearrangement, exchange and elimination reactions though these terms are used in a very broad sense. Furthur subdivision is possible on the basis of the reagents but this is not included in the symbol string. When a reaction involves more than one bond change, multiple entries are supposed to be made although this does not always appear to occur(69) and one also finds that the set of reactions denoted by a single symbol string often bear little relationship to one another(26): both of

these points are discussed at some length by Vleduts who also points out that it is often easier to find a reaction via the subject index rather than via the bond classification(52). The French firm Roussel-Uclaf operate a card file based on bond formation and reacting group data and this method of classification has been employed in the series "Cahiers de Synthese Organique"(70). Bond change data is also included in the "Chemical Reactions Documentation Service" run by the Derwent organisation(71) but it is not clear whether the bond change indexing involves mechanistic considerations(72).

The most fruitful development of Weygand's idea has been the concept of the reaction centre, or reaction site, which seems to have been first described by Vleduts(52). In his paper, he advocated the use of all the bond changes occurring during the reaction, rather than the single changes considered by Theilheimer. As he points out "a distinctive feature of organic reactions, which involve complicated molecules containing almost exclusively covalent bonds, is the destruction and creation of a comparatively small number of bonds in such a way that, during the process, fairly extensive portions of the molecules do not change their structures". This being so, we may attempt to classify reaction information upon the basis of the bonds that have been altered in the course of the reaction; taken together, these bonds represent the partial structures involved in the change, the reaction centre. To quote again "the essence of the work in developing a skeleton scheme of a particular reaction lies in the comparison of the structure of the final and initial molecules and in discarding the fragments of the structure not undergoing changes in the course of the reaction". Such a skeletal reaction scheme will generally represent several similar reaction types since groups adjacent to the reaction sites are omitted although they may play a significant part in determining the course of the reaction in terms of yield, stereochemistry and overall structural

change. The neglect of the nonreacting parts of the molecules is claimed as an advantage(52,73) since supposedly useful analogies may be detected between different reactions belonging to the same basic class but, as we shall see later, there are no generally available guidelines as to exactly what should be included in the reaction centre. Vleduts suggested that the site should consist of all the bonds altered during the reaction plus the following:

(i) any heteroatoms that are directly connected to an atom in the reaction site(a key atom)

(ii) any atoms connected by multiple bonds to a key atom

(iii) any groups of the form A=B or A≡B where A and B are any atoms of which at least one is attached to a key atom.

Mockus has pointed out that this selection of "activating groups" is made upon a structural basis rather than upon the basis of any mechanistic considerations. Of greater importance in the context of this thesis is that such groups could be detected algorithmically with relatively little effort whereas the identification of the actual activating substructures would imply a high degree of machine intelligence and significantly greater computational requirements. A greatly extended list of features has been described by Bersohn and Esack(43).

Vleduts described a method for the unambiguous linear codification of reaction centres and stated that the resultant notations were to be used as the basis of a systematic reactions index to the Russian abstracts journal Khimiya, the chemistry section of Referativnyi Zhurnal. Mockus(36) states that no such index has actually been produced but Vleduts later described a simplified version of his approach which had been applied to a file of reactions involving organofluorine compounds(74).

Ziegler, in the"Reactiones Organicae", has produced a set of punched

cards embodying the reaction site concept(73). Each card bears a skeletal reaction scheme and the structure of the product, these being described by a simple fragmentation code, as well as a printed abstract and additional physical information such as conditions and heighbouring groups. The advantages cited by Ziegler are

(i) a precise definition of the reaction type, independent of generic types such as oxidation

(ii) easy detection of analogous reactions as only the reacting parts of the molecules are coded

(iii) no assumptions are made as to the mechanism of the reaction

(iv) the use of traditional symbols since the skeletal scheme is printed upon the card as well as being punched for machine use

(v) independent of nomenclature

(vi) easy linking of the reaction centre with the whole molecule

(vii) easy classification of reactions.

Some of these advantages seem rather trivial whilst (vi) has been shown to be incorrect(see below) and (vii) is highly optimistic but it does seem that a reaction centre approach holds distinct promise and it has played a large part in the fully automatic indexing procedures to be described in the third section of this chapter. The Pharma system, which forms the basis of Derwent's CRDS and is prepared manually(36,71), has a limited amount of reaction centre information and the experimental reaction file at ICI Pharmaceuticals, again based on computer processing of manual input, is based entirely on the reaction centre approach(37,75,76). This system has been evaluated(77) and the conclusion reached that in many cases, the reaction centre alone is insufficient to characterise the reaction which implies that information is required about the wider chemical environment of the site. Identical conclusions have been reached by Osinga in his use of 'direct annotating environment numbers'(78).

Two widely used reaction documentation services are those developed by the Internationale Dokumentationsgesellschaft für Chemie(IDC)(26,79,80) and by the Pharma Documentation Ring(26,38,71), these being two consortia of European chemical and pharmaceutical firms.

Both systems employ manually assigned fragmentation codes which are stored for subsequent machine search. In the GREMAS code of IDC(81) each carbon atom is coded by at least one term consisting of three letters and a reaction is described by pairs of these terms corresponding to the initial and final states of every functional carbon atom modified in the course of the reaction. Various subsidiary terms are used to indicate the general type of the reaction, e.g. chain elongation or ring closure, and a variety of search techniques is available. The Pharma service, which is now marketed by the Derwent organisation, is based on the fragmentation code of Derwent's RINGDOC patent alerting service and a limited amount of bond change and condition data are also included. Other methods of reaction indexing have been described in the literature but the majority are of limited historical interest(82,83,84,85).

The great triumph of physical organic chemistry over the last thirty years or so has been the development of mechanistic theory by which it is possible to rationalise a large measure of known reactions upon the basis of inter- and intra-molecular electronic effects(86). Presumably a comparable degree of coverage could be achieved in the documentation area by employing some sort of mechanism based indexing. Qualitative descriptions of reaction mechanisms have been suggested (87,88,89,90) but a quantitative description could only be achieved by wave mechanics equations, these being pictorially represented on a reaction diagram by electron shifts, charge transfer complexes and the like.

Despite the variety of approaches outlined above, it seems not unlikely that most retrospective searches for reaction information are currently made via the CAS Subject and Substances indexes whilst current awareness facilities are provided by CAC.

I.3 Automatic methods for reaction indexing

The earliest suggestion that reaction analysis could be carried out automatically was made by Vleduts(52) and shortly afterwards, Mischenko et al. described an algorithm by which this might be performed (91). The underlying assumption was made that the bonds formed in the reaction would be different from those destroyed; thus a simple comparison of the bonds in the reactant and product molecules would reveal those that had changed. The input to the program consisted of the redundant connection tables of the reacting molecules and these were used to generate the lists of reactant and product bonds, the bond representatives consisting of the component atoms plus the bond order i.e. simple pairs in the terminology of (92). Bonds common to the two sides of the equation were deleted and the remaining bonds were used as the basis for binary descriptors in a punched card retrieval system. Analyses were produced for 85% of a sample file of ten thousand reactions and of those analysed, circa 75% were judged as being correct. Such a method of analysis will be at fault if some of the bonds that have been broken are identical with some of those that have been formed since the procedure would have registered them as having played no part in the reaction; such incorrect equivalences might have been expected to be quite frequent due to the very small size of the fragments used but, in fact, the results are very encouraging when one considers the simplicity of the procedure. The authors state that the detection of the changed bonds should be a mere precursor to the automatic production of a skeletal reaction scheme but this does not seem to have been carried out(93). It is perhaps worth mentioning at this point that this simple algorithm has formed the basis of much automatic analysis research to date; this includes the program to be described in the second chapter of this dissertation.

An alternative to the direct identification of structural

differences is their detection as a result of the identification of structural similarities and this was first attempted by Armitage and Lynch(69,94,95,96), similarity being defined as the largest connected set of atoms and bonds common to the structures on the two sides of the reaction equation. The method was based on the generation of fragments of each structure, starting with the individual atoms of each, and, by concatenation, fragments of increasing size. At each step in the process, the fragments formed from one structure were compared with those from the other, non-common items discarded, and growth continued in the subsequent iteration only from those fragments which were common to both. The procedure terminated when the structural 'highest common factor', i.e. the largest connected set of atoms and bonds, had been identified. Most of the work concentrated on acyclic structures where the building blocks of the common structure were linear chains of atoms. Once these substructures could be grown no further, the maximal common substructure was obtained by joining the straight chains together, thus allowing the identification of branched substructures(96).

Although intuitively appealing in that the procedure to some extent mimiced the mental processes of a chemist who, upon scanning an equation, identifies the common features as a preliminary to pinpointing the differences,it was found that the complexity of the programs became quite unmanageable for all but the simplest molecules since the number of chains that needed to be considered rapidly became very large. This was partially alleviated by pruning those smaller chains which were completely contained in larger ones but even with this modification it was found that one of the examples, in a sample file of 22 reactions from CAC, produced over 80 common chains of varying sizes. A recent development has been

reported by Cone et al(97) who found that the identification of all common substructures larger than some threshold size required upto a hundred seconds of CPU time per reaction so that it would seem that other approaches must be considered if the detection of similarities is to be of use in a practical environment where many thousands of reactions need to be processed.

By far the most sophisticated application of Vleduts' original suggestion has been described by Harrison and Lynch(98). The differences between the reactant and product structures were again determined by analysing the two sets of reacting molecules into small bond centered fragments; a variety of fragment types were investigated but most work employed the bonded pair(92). After fragmentation, the two sets of pairs were compared with regard to both the type and to the number of occurrences of that type. The analysis yielded three fragment sets, any or all of which could be empty, for each side of the reaction equation. The common pairs represented features that were unchanged in the course of the reaction and could be eliminated; the extra pairs corresponded to types which contained a greater number of pairs on one side of the equation than on the other whilst the non-common pairs represented types which were not present at all on one side of the equation.

If the pair analysis was successful in detecting some structural differences the fragments were joined together to form a skeletal reaction scheme. The assembly of the reaction sites, which was carried out separately for each half of the equation, was essentially the construction of a partial connection table record of the reaction, the assembly being carried out in much the same way as one might construct a jigsaw puzzle with the parent tables acting as a sort of template for the rebuilding. The non-common pair types were specifically defined so these caused no problems when the reaction

sites were reassembled. Where extra pairs were concerned however, a number of the pairs needed to be selected from the total of this type. The algorithm was constructed so that it would choose those extra pairs which, for the given sets) of non-common pairs, would yield the most compact reaction site(s) if an alternative were possible. Once the sites had been generated and validated, they, or rather the partial connection tables that represented them, were compacted for storage and written to an output file; in effect this meant that the nodes in the site were renumbered and correspondence with the parent molecule lost. As the OSTI report points out(99), it would be possible to form reaction files directly from the pair analyses but in this form it would not be possible to investigate the pair interconnections; furthermore, rebuilding the analysis fragments into partial reaction sites was found useful in validating the analyses.

To test the worth of an automatic analysis program, three features need to be evaluated:

(i) the percentage of the file analysed

(ii) the percentage of (i) which is correctly analysed

(iii) the usefulness of the analyses in a retrieval system.

All three were investigated(49,77,100). It was found that analyses were produced for between 79 and 97% of the reactions in a variety of files; of these circa 95% were analysed correctly, that is to say in an intuitively reasonable manner, the failures occurring primarily in reactions where extra pairs played a large part in the rebuilding of the reaction site. However severe deficiencies were revealed when test searches were carried out on the residue files which contained the partial reaction sites with simple molecular formula and bond count screens; if a query passed these screens, the search program then attempted to match the query and reaction site bonded pair sets. Queries were run against a file of 1306 reactions from

CAC and retrieved a total of 582 reactions, 53% of which were considered to be false drops(100). Queries involving acyclic features were generally reasonably effective, the failures being overwhelmingly due to the search program's inability to differentiate either the size or the nature of the monocycles in fused heterocyclic ring systems. This lack of success would seem to be primarily due to the rudimentary screening system employed; this could, of course, have been improved in the light of furthur experience. A much more serious objection is made by Seddon who, after comparing the analyses produced by three reaction centre analysis methods, states "the problem with all reaction centre analysis methods is that they may not include sufficient information about a reaction to characterise it. The technique of producing a reaction centre representation is to enable all compounds which react in a similar way to give similar reaction schemes regardless of differences in the environment of the reaction site. But here the aim of the complete retrieval system must be considered; whether broad classes are required or detailed description to answer specific queries. For the latter some indication of the environment may be essential"(77). Campbell(100) and Clews(49) found that varying the level of query specificity proved helpful but the latter concluded that the most useful approach to the problem was one where searches could be carried out both upon the analyses and upon the parent molecules, the reaction sites merely being represented by an indication of the appropriate atoms in the parent connection table.

While this work was being carried out, a study was also being made of WLN records of chemical reactions(101,102). There are disadvantages with linear notations due to the lack of explicit connectivity information and to the fact that a few WLN symbols may represent quite large numbers of atoms and bonds, which implies

that in some cases one will only be able to describe the changes in rather broad terms. More importantly, the assumption is made that there will be a close correspondence between the WLN symbol changes and the structural changes that they are meant to represent; this assumption will not be justified in all cases.

The advantages of a WLN based system are threefold:
(i) as the symbols provide printable character representations of the structural features involved in the change, one may think in terms of printed indexes of reactions similar to KWIC and KWOC compound indexes(this was the original starting point for the work though none were, in fact, prepared)
(ii) the notation gives especial prominence to ring systems and to functional groups: simple programs will hence be sufficient to handle these synthetically important features. Also, a manual assessment of part of the file indicated that a considerable portion consisted of either ring cleavage or conversion reactions and acyclic functional group interconversions.
(iii) many organisations have WLN structure files so that any results achieved might be of quite general interest and utility.

The data base for this work, which has been used for all subsequent research in this department, was constructed from the ISI publication CAC. All new compounds recorded therein are allocated an Index Chemicus Registry System (ICRS) number and the structures encoded as WLNs which are available on magnetic tape. Ten months issues of the hard copy version were scanned manually, the ICRS numbers corresponding to the reactants and products associated with each reaction being selected; these numbers were then written to a tape which was used to retrieve the required WLNs from the compound tapes which had been kindly supplied by ISI. A full description of this procedure, together with the subsequent

conversion of the WLNs to connection table format, is given in (49).

As a first step, a program was written to determine the differences in the non-numeric WLN symbol counts between the reactants and the products, these differences being assumed to have been engendered by the reaction. It was found that over 60% of the analyses were unique whilst a few analyses occurred many times; thus a fairly small dictionary of symbol changes would prove sufficient to characterise a significant percentage of the file. The use of single symbols was obviously restrictive and so the dictionaries consisted of the WLN symbol strings of the reactant and product groups involved in 41 simple functional group interconversions; these dictionaries were then searched for the corresponding symbol string changes in the reacting molecules. The identification was checked by comparing the molecular formula change calculated from the WLNs with that to be expected if the group change had in fact taken place.

It was found that 19.5% of the reactions in the file were analysed by the routine and manual checking showed that the vast majority of these had been correctly processed. However the quality of the analyses was very variable since the immediate environment of the reaction site would not be adequately described if fairly specific queries were to be expected. It would presumably have been possible to include the symbol strings of the immediate surroundings of the changed groups and then to use a longest match routine at search time; however the distribution of reaction types mentioned above means that a very large dictionary would have been required to increase significantly the scope of this approach. We may also note that problems would be expected to arise if anything but very simple groups were considered since the WLN ordering rules(17,18) might cause the symbol strings to be split(103).

Also, reactions in which different numbers of carbon atoms were gained or lost would all be identified as being of the same type since only the functionality changes were considered. The percentage of reactions analysed was relatively constant over a number of files from the same source, the actual proportion being strongly dependent on the source of the data. Thus, any dictionary would be of rather limited application outside of its source file. Seddon (77) carried out some limited trials using only the changes in specified WLN characters and symbol strings without any dictionaries to relate these to a reaction; however she gives little detail of this work.

A study was also made of the utility of WLN records in delineating reactions in which changes occur in the ring systems of the reacting molecules(102). WLN describes complete ring systems so that it was necessary to develop routines to describe the individual monocycles present; the ring lists for the reactants and products could then be compared to identify any differences caused by the reaction. The program considered only the ring changes and thus any simultaneous acyclic changes were ignored; the changes were also limited to a single monocycle on each side of the equation. 22% of the reactions in the file were processed and of these nearly 98% were subsequently judged as having been correctly analysed.

Taken together, the two routines could be expected to produce analyses for 41.5% of the file but there was little or no prospect of increasing this figure. Although the undetected failure rate was very low we should emphasise the rather crude nature of the analyses; neither program gave any consideration to the environment of the changes that it had detected and the functional group interconversion segment ignored alterations in the basic carbon

skeleton in much the same way as the ring segment ignored simultaneous acyclic changes.

Clinging suggested that the main value of the work might prove to be as some sort of screening system and this idea was extended by Lynch et al(40,104) and Nunn(105). The hyperbolic distribution of reaction types noted above was again used as a starting point, the main aim of the work being to characterise a high percentage of the reaction types in the file by relatively simple algorithms so that only those that remained would need to be submitted to the reaction site program described in (98). Interest was concentrated on reaction subfiles representing the same molecular formula change, rather than on changes in the WLN symbol counts, and it was found that circa 40% of the file comprised reactions that could be characterised by changes such as $\pm H_2O$ and $\pm H_2$. Although it was clear that these were gross changes which might disguise more complex reactions, an examination of a sample file showed that the simple molecular formula changes very often gave a correct analysis.

The adopted procedure was based on a series of three screens. The first was the molecular formula change which narrowed down the number of possible skeletal reaction types quite considerably and each of these screens was associated with a set of secondary screens which analysed the types within the subfile by looking for the presence of certain non-common and extra pairs in the reactant and product connection tables. The third screen acted as a final check and consisted of making modifications to the connection table of one of the reacting molecules in order to simulate the table of the other molecule in the reaction.

The analyses were compared with those described by Clinging and Lynch(101) and it was found that a similar group of reactions

was being dealt with although the consistency of the analyses was somewhat greater. As each subfile was considered, it became obvious that greater effort was required to analyse the increasing variety of types present in the subfiles. It was also found that the relatively simple algorithms that had been developed were inadequate for analysing reactions involving changes in ring systems(28), and it was hence decided to reconsider the use of WLN for analysing such reaction types.

The approach involved the comparison of WLN symbol strings in the reactant and product molecules as the means of identifying the reaction site but, in contrast to the previous work, the method was not limited by the use of any kind of dictionary. The file was organised into categories according to the type of ring system present to facilitate subsequent processing and three main classes of reaction were identified, these being
(i) reactions with no apparent change in the numbers or sizes of rings
(ii) reactions with a change in the numbers or sizes of rings
(iii) other types, these including acyclic reactions and those involving molecules containing benzene rings only.
Programs were written to analyse the first two types which comprised circa 80% of all the reactions in the data base.

The first class of reactions consisted mainly of changes in the acyclic components of cyclic molecules although there were sometimes minor changes both inside and outside the ring brackets (the WLN symbols L,T and J). The analysis consisted of comparing the ring substituents one locant position at a time, thus allowing of changes at more than one substituent position. Checks were also made for certain symbol interconversions within the ring brackets to cover such elementary reaction types as reduction of ring carbonyls and the hydrogenation of unsaturated linkages. The

second class of reactions was analysed in two stages. The first of these was to identify the changes in the ring system which enabled summaries of ring changes to be produced, the procedure being based upon Clinging's algorithm. The second stage of the analysis was to determine additional changes other than those occurring in the ring systems, this being carried out using the algorithm designed for the first class of reactions; it was hence possible to provide descriptions of all the parts of the molecules that had been involved in the reaction. No algorithms were developed for the third class of reaction types although it was claimed that a procedure analogous to that used for the first class could be employed; in toto, circa 70% of the reactions in the file were analysed. A trial index for the first class was produced in which the sort key was the WLN symbol strings produced by the analysis.

Although satisfactory in many respects, this work suffers from several deficiencies. Firstly, the ring change algorithm was not very specific in that there was no way of connecting the intra ring changes with any simultaneous substituent changes. Next, in many of the reactions, the entire substituent WLN strings were given as the analysis even though large sections of them may have been unchanged i.e. the exact site of the reaction was not specifically defined. Also problems might arise due to the ordering rules of the notation.

Osinga and Stuart(39) have described a faceted classification scheme(106) for reactions which contains eight main facets, these including addition, elimination and ring changes. The aim of their research is claimed to be the automatic classification of reactions using this scheme but it seems to have been applied only to a file of seven reactions, one of which was incorrectly processed(107).

WLN was used as the input structure representation for the reacting molecules and the WLNs were used to produce a sort of connection table in which the nodes were described by direct environment annotating numbers(DEAN), these being integers representing atom centred fragments similar to augmented atoms(78,108). The approach used in the generation of the connection tables was presumably similar to that adopted by Hyde et al.in their work on the CROSSBOW project(109,110) and by Granito et al.during the CHEMTRAN project(111,112,113); it seems that the DEANs play a similar role to the 'units' section of a CROSSBOW table(16). Reactions were analysed by generating lists of augmented bonds from the connection tables of the reacting species and then deleting those common to both sides of the equation i.e. the approach was basically that of Mischenko et al(91). The assignment process by which the analyses were used to produce the appropriate classification is unclear.

All of the procedures outlined above involve some form of fragmentation which implies that a degree of ambiguity may be present in that it might not be possible to determine the exact location of the reaction sites in their parent molecules. Vleduts has suggested a method for indexing reactions which, potentially at least, could overcome this problem. He stated that "the ultimate objectives of the algorithmic analysis of reactions in the framework of the approach is the detection of the exact locations of the chemical bonds altered by the reaction...and the identification of the exact nature of the changes involved"(114). His approach consisted of an atom by atom mapping of one reacting molecule onto another to identify the largest common substructures and, in consequence, the differences engendered by the reaction. Once the atoms in the common substructures had been mapped it would be a simple matter to identify the bond changes that had occurred.

The algorithm involved the identification of the maximal subgraphs common to the two sides of the reaction equation; in contrast to the problem of graph isomorphism(115,116,117,118,119), maximal subgraph isomorphism has been little studied due to the greater complexity of the problem(120,121,122). It is well known that isomorphism may be determined by a simple enumeration algorithm(119); in the present context a possible procedure would consist of generating all possible subgraphs(partial structures) from one graph(reacting molecule) and then matching them against all possible subgraphs from the other(120). The computation required may be substantially reduced if properties of such subgraphs which are invariant under isomorphism are taken into account; thus a reactant atom may not be mapped onto a product atom if the atom types are different. Such 'set reduction' techniques, initially described by Unger and Sussenguth(115,116), form the basis of the iterative search procedures used in searching structure files(123,124). A method for maximal subgraph identification will only be of practical utility if the process of subgraph matching is simple and efficient; indeed Vleduts suggested that such an algorithm, implemented upon current hardware, would probably be limited to structures not exceeding ten to fifteen atoms. He accordingly described a procedure whereby a comparison of the WLN symbol strings of the reacting molecules would be used to provide possible reactant-product atom equivalences to reduce the amount of iterative mapping that would need to be performed. Neither stage of the procedure was implemented.

CHAPTER II

The use of Wiswesser Line Notation records in the automatic analysis of chemical reaction data.

II.1 Introduction

In the previous chapter, we have given a detailed account of the work carried out in this department on the automatic indexing of chemical reactions. Two distinct approaches to the problem were identified. Firstly, an attempt was made to map the structures of the reactant and product molecules onto one another so as to identify the largest common substructures and hence the differences by subtraction(94,95). The work was abandoned owing to program complexity and the amount of processing time required. More recently, Vleduts described an alternative algorithm by which the mapping could be achieved but no attempt to implement the procedure was forthcoming(114). We shall return to this approach in the following chapter of this thesis. The second approach involved a comparison of the reactants and products to identify the differences directly: both connection tables and Wiswesser Line Notations were used as the structure representation(40,98,101,102).

The earlier work involved a whole structure fragmentation process in which the redundant connection tables of the reacting molecules were broken down into sets of small molecular fragments called bonded pairs. It should be noted that in the chosen fragmentation mode, all parts of the molecule were described in equal terms: this is in direct contrast to the majority of the fragmentation codes used in screening systems which consist of a limited number of chemically significant features which are specifically searched for in the structure to be screened(125). Although the use of a whole structure fragmentation ensures that all the fragments present can be described in some way, it does mean that the chemically significant features, such as functional groups and rings, may not be sufficiently delineated. Also, the fragments used were quite small, two atoms and the bonds around and

between them, and thus of relatively high frequency and of low variety, i. e. the sets of fragments produced by the analysis often consisted of several occurrences of a limited number of fragment types. This led to severe problems when the non-common fragments were reassembled to produce a skeletal reaction scheme(99). It was also found that the more specific the fragment type, the better the chance that a successful analysis would be achieved(98): thus larger fragments such as bonded pairs and augmented atoms were found to produce better results than fragments such as augmented bonds and bonded atoms(92,108).

The early WLN work considered only a limited number of structural features, specifically monocycles and some of the more common functional groups, in the reacting molecules(101,102). The procedures could only detect certain of the changes taking place since some portions of the reacting molecules were not considered. Further work(40,104) concentrated on reactions involving molecules containing ring systems, these being found to represent over 80% of the reactions in the file used for the study. All parts of the reacting molecules were considered and thus reactions involving both cyclic and substituent changes could be analysed as such although no attempt was made to link the changes together to form a reaction site. Also, the substituent changes were described in rather generalised terms so that it was not always possible to locate the changes within the molecules. Finally, the method of ring analysis was limited in scope since it relied, in part, on a dictionary of ring heteroatom symbol changes.

The work described below represents an attempt to combine features of both the connection table and WLN methods of reaction analysis: in particular, an algorithmic structure fragmentation procedure is adopted, together with a residual fragment rebuilding routine to

produce a reaction site, but this is designed to be applied to WLN structure records. Analogous fragmentation procedures have been described by Bawden(126) and by Hyde et al(110) but the method developed here would seem to be considerably more detailed since provision is made for the generation of descriptors at four levels of detail, these levels being chosen on the basis both of a knowledge of the reaction types in the Shefield file and by the way in which WLN delineates the various kinds of substructural feature that may be expected. The procedures have been chosen so as to produce fragments representing chemically significant groupings and thus they may be expected to describe adequately common reaction types such as functionality changes, elimination and ring conversions with the minimum of processing: at the same time, provision is made to allow of a description of all possible types of reaction.

The rationale for a multilevel fragmentation is simple, but does not appear to have been explicitly stated in the context of reaction analysis. Algorithmic fragment generation is routinely used as a means of obtaining potential screens for searching chemical structure files: as will be discussed in the fourth chapter of this thesis, efficient screen sets are obtained by consideration of the distribution of fragment incidences, the fragments in the screen set being chosen so that each screen occurs approximately equifrequently across the whole file. The procedure described here is based on a very different criterion since we wish to produce fragments which are as large as possible, subject to the constraint that they represent features common to both sides of the reaction equation. Once these large, common features have been identified they may be discarded and a more specific fragmentation method adopted to remove furthur common features. The process continues

until, hopefully, the remaining truncated structures represent the reaction sites. This simple principle is the basis both for the method of WLN analysis described below and for the approximate structure matching algorithm presented in the following chapter.

II.2 A multilevel WLN fragmentation procedure

In this section we present the four fragmentation procedures that have been developed to decompose a WLN symbol string into a set of substructural representatives. The description is mainly by example since some of the details are rather complex.

From a consideration of the ring size numerals for the reactant and product notations in a file of 9197 one reactant/one product reactions, Lynch et al.(40) found that circa 50% were reactions in which no change was apparent in the number or sizes of the ring systems or the molecules involved. It should be noted that this figure does not include molecules containing only benzene rings and does include reactions in which there were changes in certain of the ring heteroatom symbols(104). Nevertheless, it would seem that in a large number of cases, the basic ring systems remain unaltered, the changes being confined to the substituents: note that in our work, phenyl groups, denoted by the WLN symbol R, are considered as ring systems in just the same way as those described within the ring delimiters L, T and J. This being so, three types of feature are considered in the first level of fragmentation, these being ring systems, phenyl groups and ring substituents. The two notations are scanned, symbol by symbol, and any such groupings are noted and stored, acyclic molecules having been isolated previously. An example of the method of analysis is shown in Fig. II.1 for which reaction we obtain the two fragment lists

(1) L E5 B666 BUTJ, *1, *1, *Y1&UNZ and

(2) L E5 B666 BUTJ, *1, *1, *Y1&UNNUY1&1

where * represents an attachment to a ring of some kind. Elimination of the common fragments then yields the analysis shown in the lower

half of the Figure. Two further examples are shown in Figures II.2 and II.3. One should note that in all cases, the fragments resulting from the analysis are passed on for further processing so that we may delineate the reaction site as precisely as possible; only for the reaction of Fig. II.3 would the final analysis be the same as that provided by this level of fragmentation.

More generally, there will also be changes in the ring systems of the reacting molecules so the second fragmentation involves a description of any ring systems in the reacting molecules which have not been eliminated by the first level analysis: as in the previous work carried out in Sheffield, the ring systems are described in terms of their constituent monocycles. It is worth considering exactly which monocycles should be considered since even a simple system such as decalin may be thought of as consisting of either two fused six rings or a single, bridged ten ring whilst for more complex systems, the number of possibilities proliferates e. g. cubane contains a total of twenty eight rings of various kinds. The detection of all the possible rings in a compound or, more generally, the circuits in a graph, has been studied by many workers while other investigations have concentrated on some limited subset of the potential ring set: for further details the reader is referred to a paper by Wipke and Dyott(127) which discusses over a dozen different ring perception algorithms. More recently, Zamora has given an algorithm for detecting the smallest set of smallest rings(128) and this subset is the one used in the present work since it is these rings which are explicitly defined by the WLN of a compound.

Having decided which rings are to be investigated, we must identify the constituent atoms of these rings so that they may be characterised in terms of, e. g. size and heteroatomic character;

finally, the exact level of description must be selected. A simple algorithm for detecting the constituent atoms of each monocycle in a WLN has been given by Granito *et al*(112) and by Palmer(132): Clinging's adaption of the former procedure yields a canonical description of the atoms in each of the rings(102). However, one of the primary objectives of the work is to generate reaction descriptions which could be used in a printed index and it was felt that the descriptions provided by Clinging's algorithm, a list of the atom types together with their degrees of saturation, would neither be simple to scan nor compatible with the rest of the index entries which were to consist of WLN symbol strings. It was therefore decided to encode the constituent atoms in a form of simplified WLN using the WLN ordering rules for monocycles(17): a detailed description of the way in which this was achieved is given in Appendix I. Eakins(129) and Adamson *et al*(130) have described a range of possible levels of ring description but none of these are directly applicable here due to the use of WLN, rather than of a connection table, as the structure representation: thus the WLN symbols Y, V and S*W*, when contained within the ring delimiters L, T and J, all represent extra-ring attachments of some kind whilst many of the attachments are not described explicitly and may only be detected by a consideration of the ring substituents. Problems also arise when one comes to assign a ring saturation symbol, T or &, since in a complex fused system, it is often dificult to state the exact degree of saturation of each atom and/or monocycle.

The descriptions provided are, perhaps, best illustrated by example. If we consider the reaction shown in Fig. II.4, the fragment sets produced after the first level analysis are

(1) T66 BMT&J and (2) T66 BNJ,

the benzene rings and their substituents having been eliminated.

The ring analysis algorithm yields the descriptions

(1) @L6J, @T6 AMTJ and (2) @L6J, @T6 ANJ

where @L6J represents a fused, carbocyclic, unsaturated, six ring i. e. a fused benzenoid monocycle. A comparison of the two fragment lists gives the reaction analysis shown in the lower half of the Figure. Further examples of the method of analysis are given in Figs. II.5 and II.6. Fused rings are denoted by the prefixed symbol @ but no attempt has been made to identify the exact mode of inter-ring attachment: thus both of the ring systems shown in Fig. II.7 will yield the same fragment lists, i. e. two @L6TJ rings. For the file studied, this has not proved to be a great limitation but means are available for providing this information if it were thought necessary(131).

So far, no attention has been paid to the acyclic portions of the reacting molecules: these are analysed in the third and fourth levels of fragmentation. The third level involves rupturing the WLN symbol string whenever a terminal atom or branching symbol is identified, branching symbols being defined as any symbol that disturbs the linearity of the character string: thus X, Y and -SI- are considered as branching symbols whereas V, W, when attached to S, or N, when attached to U, are not. This method of fragmentation was chosen for three reasons:

(1) as the resultant fragments are linear, it is easy to obtain canonical descriptions for them by a simple alphanumeric comparison of the fragment character string, as generated, and its reverse. That sorting lower may then be arbitrarily chosen as the representative so that, e.g., the substructure shown in Fig. II.8 will always be described by the fragment symbol string /10V2U1/ where / represents an attachment to an acyclic branching symbol;

(2) a high percentage of functional groups remains intact under this type of fragmentation where we use functional group to describe any string of hetero- and/or unsaturated atoms: Vleduts(52), Hendrickson(34) and Seddon(77) have all emphasised the importance of such features. Clinging and Lynch(101) showed that circa 20% of the reactions in the file could be analysed using a small dictionary of functional group interconversions and the more general description used here implies that a significantly larger percentage of the file should be effectively characterised.

(3) fragmentation at the branching points of the notation allows a relatively simple reassembly of the noncommon fragments into a reaction site(see below).

Again, the fragmentation is described by example. The uncanonicalised fragment lists for the reaction shown in Fig. II.9 are

(1) QMU1/, X, /1, /1, /V5U1, Y, /1, /1 and

(2) NC/, X, /1, /1, /V5U1/, Y, /1, /1.

Elimination of the common fragments yields the analysis shown in the lower half of the Figure. Further examples are given in Figs. II.10 and II.11.

Although effective in many cases, this method of fragmentation does mean that long, unbranched carbon chains will remain intact and the fourth and final level of analysis involves the truncation of any such features present to a fixed length of one methylene unit, i. e. /1/. Figure II.12 illustrates a typical acyclic reaction together with the fragments obtained after the first three levels of analysis. Although the long carbon chain is important in describing the exact environment of the group that has changed, it does mean that in a printed index, there would be an inevitable scattering of the entries describing the hydrolysis of unsaturated acid esters because of the various possible lengths

of the methylene chains. The final fragmentation truncates the symbol strings to yield the much more general analysis shown in the lower half of the Figure, the four /1/ units having been eliminated since they are common to both sides of the equation.

II.3 A program for automatic chemical reaction analysis

The fragmentation procedure described above has been implemented in a program to produce reaction descriptors automatically using as input the WLNs of the reactant and product molecules. It will have been noticed that all the reactions illustrated so far have yielded only a single analysis fragment on each side of the reaction equation. Many reactions, however, produce several fragments and it would clearly be of value if these species could be recombined to produce a notation string describing the reaction site. Such an approach forms the basis of the work described in (98) and we have developed an analogous synthesis segment to produce a WLN symbol string characterisation of the reaction sites. In principle, synthesis could take place after each and every level of fragmentation but in practice we have included synthesis routines only after the second and third levels. In the first case, the fragments to be considered are monocycles and ring substituents, whilst in the second they are branching symbols and the pendant linear chains: as reactions involving ring systems predominate in our file(40), the first of these routines is much the more heavily used.

The choice of level at which synthesis takes place has been largely conditioned by the ease with which the requisite connectivity information can be obtained from the input notation. As WLN is a whole structure representation it is possible to generate a full atom adjacency matrix description for a large percentage of notations(110,112,133,134) but at this stage of development it was decided that the incorporation of a full connection table generation segment would be uneconomic in terms of programming effort. Moreover, we are primarily interested not so much in the interconnections of individual atoms but of the analysis

fragments. Accordingly, instead of a full atom adjacency matrix, we have used fragment connectivity lists which are built up during the running of the program. We shall consider first the ring-substituents list. During the first level fragmentation, whilst scanning the input WLN strings for ring brackets and benzene rings, a note is made of the locant position of all substituents: at the same time, a stack is operated to keep track of all the benzene rings and later ring systems after the first so that it is possible to match all substituents with their parent ring systems. If any such systems are left after the first analysis, the monocycle generation routines produce a list of all locant positions for each monocycle so that it is also possible to match analysis substituents with their parent monocycles. During rebuilding, the analysis fragments are joined together in a linear string using the information in the connectivity lists: where a choice is possible, the program chooses the non-overlapping site or sites which are most highly connected, i. e., those comprising the largest number of analysis fragments. The resultant sites are compressed to a continuous linear symbol string for output.

The rebuilding is done in two stages: firstly, where appropriate, substituents are joined to benzene rings and then these larger fragments are joined to their parent monocycles. The procedure will be exemplified by the reaction shown in Fig. II.13. The fragment lists obtained after the first two levels of fragmentation are

(1) @T6 AN DNJ, @L6J, *Q, *OSW*, R and

(2) @T6 AM DMJ, @L6 AV DVJ.

Inspection of the connectivity records for the two reactant phenyl groups, the WLN symbol R, shows that one of them is attached to an

OSW grouping and it is accordingly assumed that it is this phenyl group that has been involved in the reaction. Merging the two fragment character strings yields the string *OSW* R. A subsequent inspection of the four @L6J reactant fragments' connectivity lists shows that one of them has both an *OSW* and a *Q substituent attached to it: the character strings are hence merged to yield @L6J *Q *OSW* R. No analysis fragments are found to be attached to either the reactant @T6 AN DNJ or the product @T6 AM DMJ and @L6 AV DVJ rings and so the procedure terminates with the analysis, i. e. reaction site notations, shown in Fig. II.14. No details are included as to the manner in which individual monocycles are joined together so that if two analysis rings were fused, the fact could only be noted by an inspection of the parent ring system WLN; many of the ring change reactions in our file are found to be confined to a single monocycle so that we do not consider this to be a major problem. It should also be noted that the exact substituent ring positions are not specified so that we are dealing with a form of Markush structure; however, the trial searches carried out to date suggest that this is not a serious omission.

The second set of synthesis routines is used for acyclic molecules after the third level fragmentation and analysis. During this fragmentation a record is generated, similar to that above, but rather than noting substituents attached to monocycles we list linear chains attached to branching symbols. As an acyclic molecule can be considered as a tree, it is relatively simple to reconnect all the fragments so that we have a true whole structure representation whereas the ring-substituent synthesis routines produce a much more generalised description of the parent molecule. The acyclic synthesis routines are exemplified by the

hydrolysis and decarboxylation reaction shown in Fig. II.15. The canonicalised fragment lists obtained after the third level fragmentation are

(1) /V02, X, /V02 and (2) /VQ, Y.

The connectivity list for the first reactant analysis fragment, /V02, shows that its only attachment is to the second fragment, X, which also appears in the analysis fragment list: the fragment character strings are hence joined to give the string 20V/ X. The attachments of the second fragment, the tetravalent carbon, are then considered and this results in the reactant reaction site string 20V/ X /V02. Similarly, the product reaction site string is obtained as QV/ Y. The analysis and reaction site notations are shown in Fig. II.15.

Although somewhat crude in design, it is found that the two types of synthesis routine are quite effective in providing a general description of the change. Both routines operate by considering each fragment set one fragment at a time, this including both analysis and common fragments, and then commencing reaction site symbol string growth whenever a potential analysis fragment is found. This allows the selection of the largest, i. e. greatest number of included analysis fragments, reaction site if a choice is possible. Thus, for the reaction shown in Fig. II.16 the single reactant site notation /40/ X /1 /1 /1 rather than the four part /1, /1, /40/, X /1 would be chosen. After the largest single site has been obtained, the algorithm cycles back through the fragment lists to produce secondary, smaller sites if there are still analysis fragments outstanding.

Further illustrations of the synthesis routines are provided by the reactions and notations shown in Figs. II.17 to II.19.

The basic segments of the program have now all been described and an outline of the overall algorithm is as follows(note that

most of the steps are applied to both the reactant and product structures):

(1) read in reactant and product WLNs.

(2) fragment the molecule into ring systems and substituents noting their interconnections.

(3) compare fragment lists and eliminate duplicates.

(4) decompose any remaining ring systems into the component monocycles.

(5) as (3).

(6) synthesise any remaining monocycles and substituents into a reaction site.

(7) fragment molecule(if acyclic) or reaction site notation into linear chains and branching symbols.

(8) as (3).

(9) if acyclic, synthesise remaining fragments into a reaction site.

(10) truncate linear symbol chains to a length of one methylene unit.

(11) as (3).

(12) output the remaining analysis fragments, reaction sites, original notations and bibliographical citations to tape for sorting and index production.

The program, which contains about 3000 COBOL statements, was tested by processing the file of 9197 one reactant, one product reactions described earlier. The program required 587 cpu seconds inclusive of transput although this timing figure could be reduced significantly with a limited amount of reprogramming.

In all, analyses were produced for 7415 reactions, 80.6% of the file, these giving rise to 29609 index entries. The reactions not analysed can be conveniently divided into two classes: those failures arising from limitations in the algorithms and those from

limitations in the program implementing them. It was found that 1154 reactions, circa 65% of the failures, were in the latter class, these arising from a variety of restrictions such as ring systems with non-consecutive locant paths, too many or too large fragments for available storage, variable valency heteroatoms and the like. Most, if not all, of these reactions could be processed given additional programming effort. Our discussions will hence be restricted to the 628 failures arising from limitations in the algorithms used. These reactions can be divided into three classes:

(1) 404 reactions for which a unique reaction site could not be produced,

(2) 119 reactions for which no common fragments could be detected, i. e. no fragments were eliminated and

(3) 105 reactions in which all the fragments on both sides of the equation were eliminated.

Examples of type (1) failures are shown in Figs. II.20 to II.22. In the first case there are two possible reactant reaction site strings, these and the associated substructures, (a) and (b), being shown in the lower half of Fig. II.20. Since they contain the same fragments, but in different orientations, the ring synthesis algorithm is unable to prefer one possible site to the other.

A frequent reason for failure is a substituent at a fusion point since if both the rings and the substituent are involved in the change, the program cannot know to which monocycle the substituent should be allocated; this is exemplified by the methyl group in Fig. II.23, which could be attached either to the @T3 AOTJ or to the @L9 AVTJ ring.

Occasionally, an ambiguity is noted where one does not

actually exist; thus for the reaction shown in Fig. II.24, the two equally valid strings G/ X /1 and X /1 /G are produced for the product reaction site and an ambiguity is therefore presumed to exist.

Type (2) failures occur mainly with small molecules producing only a limited number of fragments such as the two reactions shown in Fig. II.25. Finally, cases where all the fragments are eliminated arise primarily from changes in ring saturation patterns since these are not explicitly defined by the monocycles produced in the second level fragmentation; the only information on monocycle unsaturation patterns is that obtained from the & or T symbols immediately prior to the J ring delimiter in the parent system WLN.

Having considered the analyses that have been rejected, what of those that have been passed as valid? Inspection of the analyses shows that reasonable descriptions are provided for a large number of the reactions processed. This is especially true for most simple functional group interconversions, many acyclic eliminations and hydrogenations and simple ring changes; in such reactions there are generally close similarities between the reactant and product WLNs and the program produces both reasonable fragments and also a valid, and useful, reaction site. Examples of such reactions together with the reaction site notations and analysis fragment lists are shown in Figs. II.26 to II.28. In cases where there is little or no similarity between the WLNs of the reacting molecules, either or both of the levels of reaction description provided may be at fault in some way. This is especially true when a ring formation occurs from a purely acyclic precursor since then the synthesis routines will be called after different levels of fragmentation and thus the reaction site

strings do not localise the change at all: examples are shown in Figs. II.29 and II.30. Minor ring changes often produce analyses of little value: thus both the analysis fragments and reaction sites do little to localise the changes involved in the reactions of Figs. II.31 and II.32.

The best test of the adequacy of the descriptions is given by actual retrieval experiments to which we now turn.

II.4 Searching a printed index of chemical reactions

The output from the analysis program has been used to produce a printed index to the file of analysed reactions. The initial mode of access is via the analysis fragments, these being the monocycles produced by the second level of fragmentation or the truncated linear chains produced in the final analysis. Having isolated potentially relevant material, the search can be made more specific by consideration of the reaction site symbol strings whilst the original WLNs are also provided as a final check. Sample pages from the index produced are shown in Fig. II.33.

It will be clear that effective searching of the index requires a fair degree of knowledge as to the methods of analysis used: however, it is felt that, given a reasonable degree of familiarity with WLN, the index should prove relatively simple to use. Trial searches were run with a set of queries kindly supplied by the staff of the Research Information Department, Pfizer(UK) Ltd. and three examples will be used to illustrate the search procedure.

The first request is for hydrogenation reactions of the form

$$R\text{-}CH{=}N\text{-}R' \longrightarrow R\text{-}CH_2NHR'.$$

The eight possible reactant analysis fragments are *1UN*, *1UN/, *1NU1*, *1UN1/, *NU1/, /1UN/, *1NU1/ and /1NU1/ although several of these fragments were not present in the index. Searching under these fragments, together with a subsequent inspection of the reaction site notations to determine the presence of the corresponding hydrogenated product character string, produced the five reactions shown in Fig. II.34.

The second query was for reactions of the form

$$R\text{-}CH(OH)CH_2OH \longrightarrow R\text{-}CHO.$$

This change corresponds to the reaction site notation

Y /1Q /Q ——→ *VH, /VH, *1VH or /1VH.

Reference to a ranked list of analysis fragment frequencies showed that the least frequent of those available for search was the reactant fragment /1Q: subsequent inspection of the reaction sites listed under this heading yielded the four reactions shown in Fig. II.35. The final query was for the reaction shown in the upper part of Fig. II.36. Searching under the reactant analysis fragments @T6 ANTJ and @T6 AMTJ produced the reaction shown in the lower half of the Figure.

It should be noted that several of the queries could not be searched in any way at all. Thus requests for condensation reactions involving diethyl phosphate and reactions involving aromatic carboxyl protecting groups are much too general while the reactions shown in Fig. II.37 and II.38 both require additional data not present in the records available. Finally, the reaction shown in Fig. II.39 would produce a quite massive output unless an initial substructure search procedure could be used to remove the vast majority of the reactions that would otherwise be retrieved. In toto, of forty queries searched, eleven retrieved some relevant material and twenty four retrieved nothing, the remaining eight queries not being searchable. The total search time was about four hours, well under ten minutes per query, though this would obviously be greater if someone other than the author were to be doing the searching.

A better test of the effectiveness of the descriptions provided has been performed in collaboration with the staff of the Research Information Department, Pfizer(UK) Ltd.. This involved a detailed comparison of the reaction descriptors provided by the WLN analysis developed here and by Derwent's Chemical Reactions Documentation Service(CRDS). Since this project was

collaborative in nature, only a synopsis of the work will be presented here: a full description is given in Appendix III which is a copy of a joint paper submitted for publication in the Journal of Chemical Information and Computer Science.

The work involved the encoding of circa 500 reactions from the CRDS data base in WLN and then producing a printed WLN index. Eighteen typical reaction queries were then searched manually by the author whilst computer searches of the CRDS descriptions, which are based on bond change information together with Ringcode(38), were carried out at Pfizer. A detailed comparison of the reactions retrieved by the two systems showed that the descriptions provided by the WLN analysis appeared to be at least as effective as those produced by Derwent's manual indexing. In several cases, the WLN results were noticeably more precise due to the range of levels of search provided but both systems were relatively ineffective for very general queries. In the case of CRDS, this was because the searches produced a very large amount of irrelevant output whilst in the WLN case, the specificity of the analyses meant that many possible search terms had to be considered. It is clear that if printed indexes are to be used in an operational situation, broader terms must be provided. For instance, the monocycles could be described simply by their size and the number of heteroatoms. Similarly, acyclic groups could be indexed at a general level by matching their character strings against those of a limited number of common functionalities such as acids, esters and nitro groups: this approach would be similar to that used by Clinging and Lynch(101). Futher details are given in Appendix III.

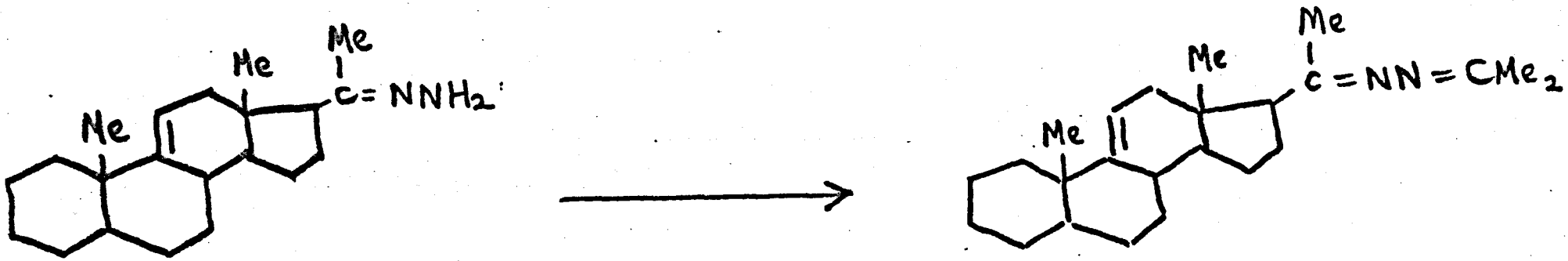
The size of the index produced could be significantly reduced if entries are not made for commonly occurring fragments:

as noted above, searching is best carried out using a ranked fragment frequency list to identify specific index terms. Thus the removal of entries under the ten most common fragments, shown in Fig. II.40, would decrease the size of the index by over a quarter with very little reduction in its effectiveness.

How discriminating are the fragments that have been produced? Fig. II.40 shows a ranked occurrence list for the ten most frequent analysis fragments, these being based on the 7415 successful analyses. It will be seen that even the most common fragment, *Q, may be expected to occur in just over 10% of the file, if the reactant or product character is stated, whilst the tenth most frequent species, *Z, occurs in less than 7% of the file in toto. Over 1500 of the 1862 different fragment types produced occur less than 10 times in the file either as reactant or product and hence the majority of the fragments are very specific in character. Entirely analogous results have been obtained by Lefkowitz in his Mechanical Chemical Code(135,136) and any algorithmic fragmentation procedure that can give rise to large fragments will invariably produce a great variety of fragment types(137,138); this will discussed in the fourth chapter of this thesis. Meanwhile, we note that if the method of analysis described here were to be applied to very large files some form of generic search capability would need to be added.

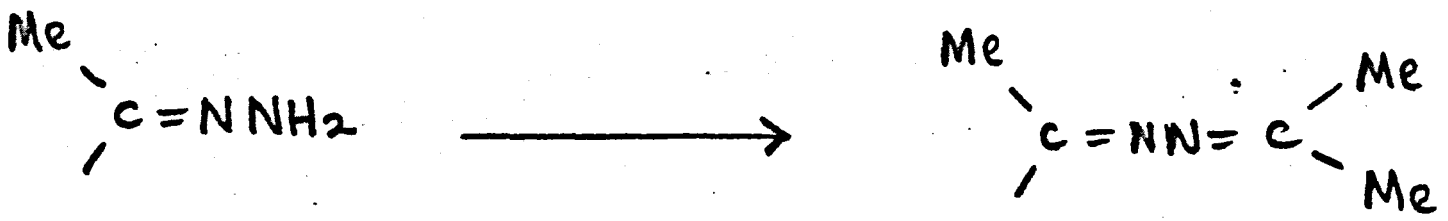
An even more disparate frequency distribution is exhibited by the reaction site notations. The ten most frequent sites are illustrated in Fig. II.41 and it will be seen that they represent very simple changes indeed. This is in line with the results of Garagnarni and Bart(139) and of Lynch(140), both of whom found that very simple changes predominated. As will be seen from the Figure, the most frequent notation occurs only 54 times and it is

found that 4452 of the notations occur once only. Such distributions are to be expected if very large fragments(which is how the reaction sites can be thought of) are considered: similar results have been obtained from files of author surnames(142). To some extent, the low frequencies are due to the fact that the reaction site notations are not canonicalised in any way: however, replacement of the reaction site strings with an alphabetically sorted list of the analysis fragments, which have been canonicalised, only increased the most frequent reaction to 70 occurrences and decreased the number of singly occurring sites to 3661.



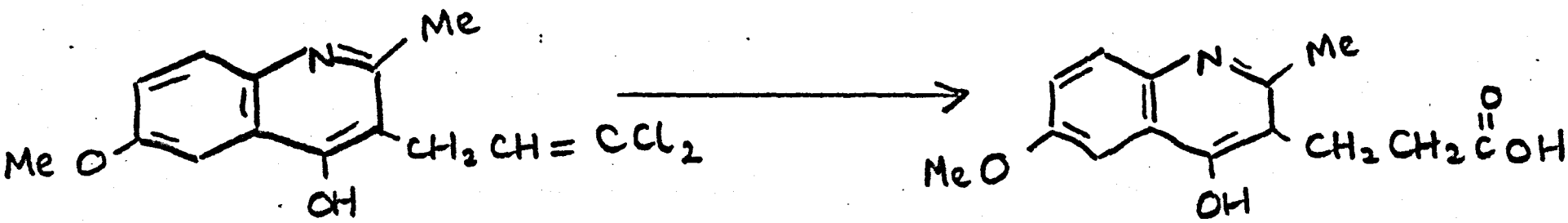
L E5 B666 BUTJ A1 E1 FY1&UNZ

L E5 B666 BUTJ A1 E1 FY1&UNNUY1&1



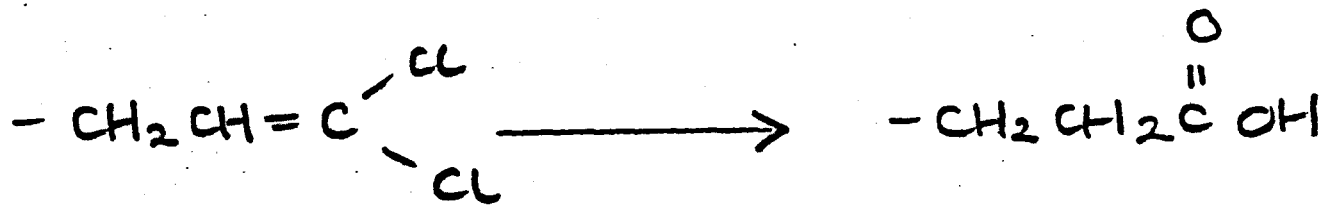
*Y1&UNZ

*Y1&UNNUY1&1



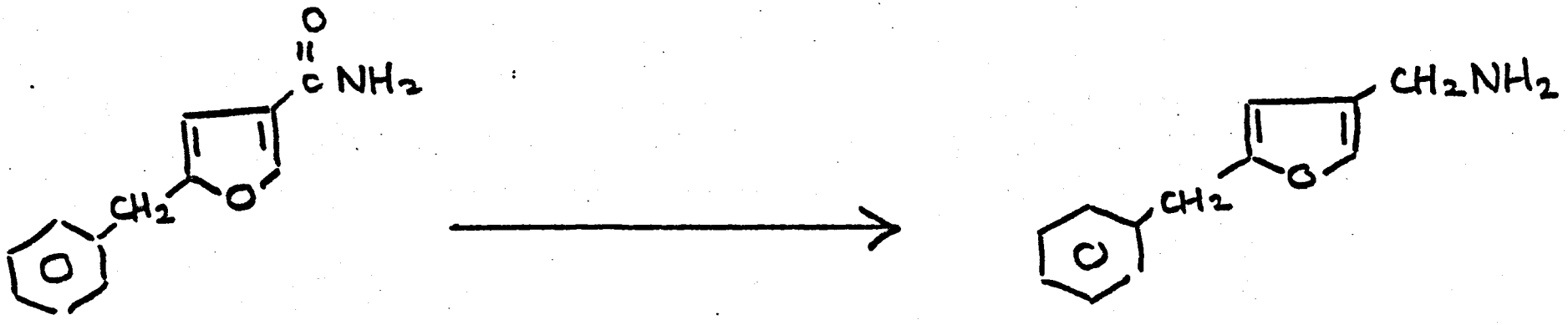
T66 BNJ C1 D2UYGG EQ HO1

T66 BNJ C1 D2VQ HO1



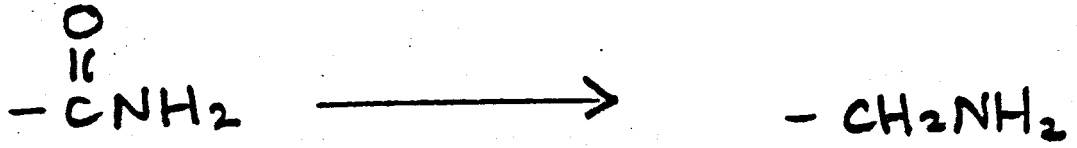
*2UYGG

*2VQ



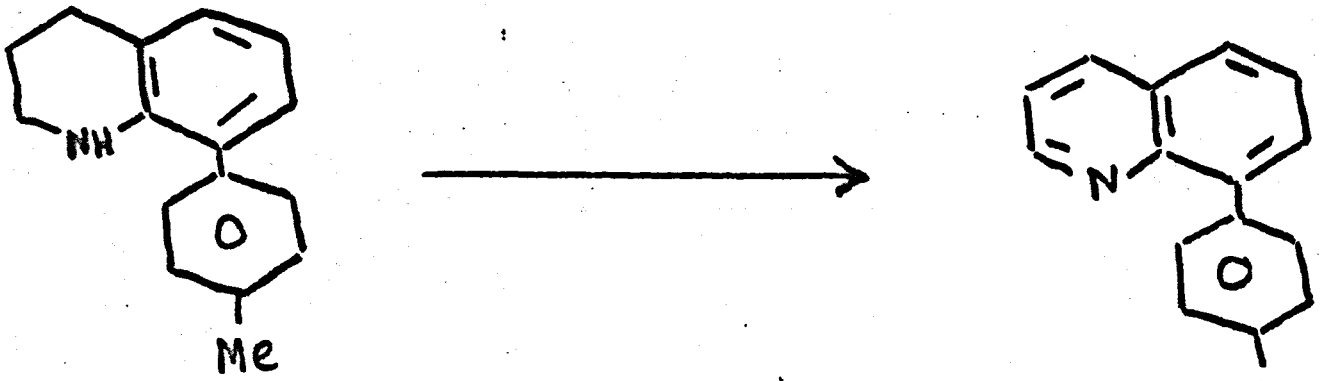
T50J B1R& DVZ

T50J B1R& D1Z



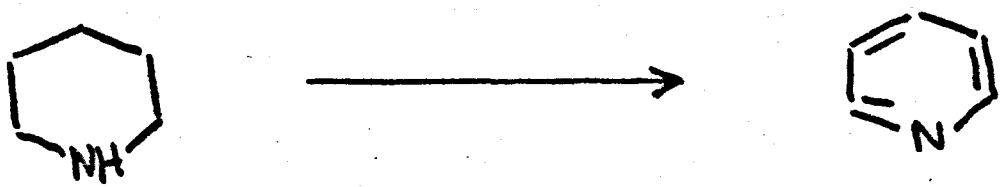
*VZ

*1Z



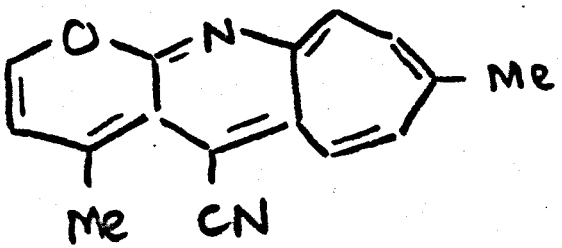
T66 BMT&J JR D1

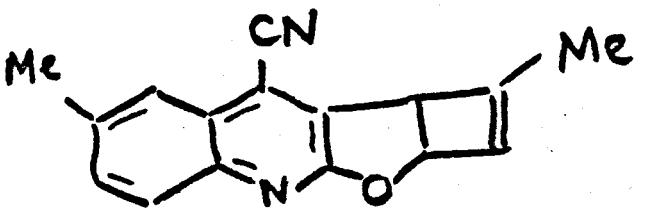
T66 BNJ JR D1



T6 AMTJ

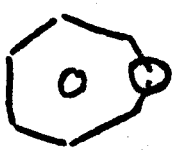
T6 ANJ





T C667 BN 00J F1 JCN L1

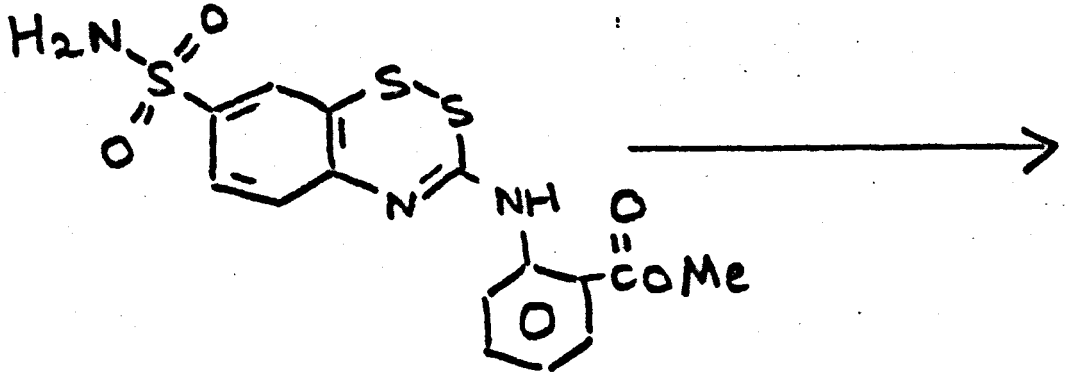
T D6 B654 JN LO NU&&TTJ CCN F1 01



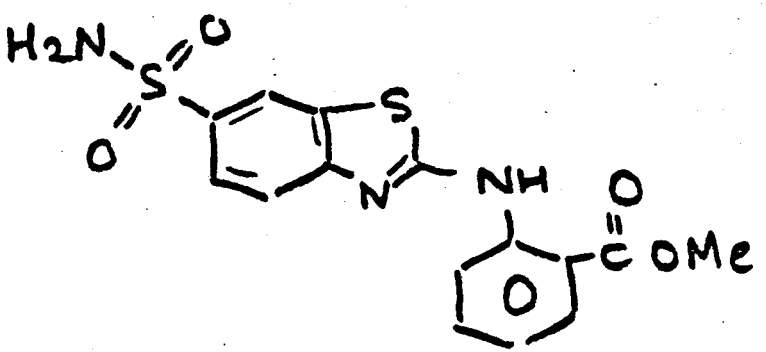
@T7 AOJ

@T5 AOTJ

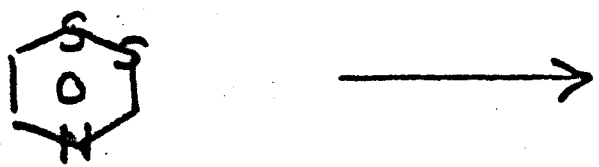
@L4UTJ



T66 BSS ENJ DMR BV01& ISWZ

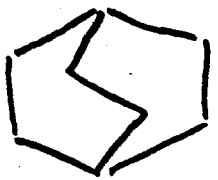


T56 BN DSJ CMR BV01& GSWZ

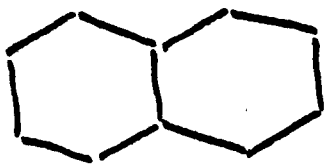


@T6 AS BS DNJ

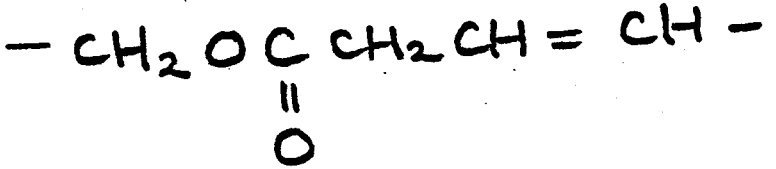
@T5 AN CSJ

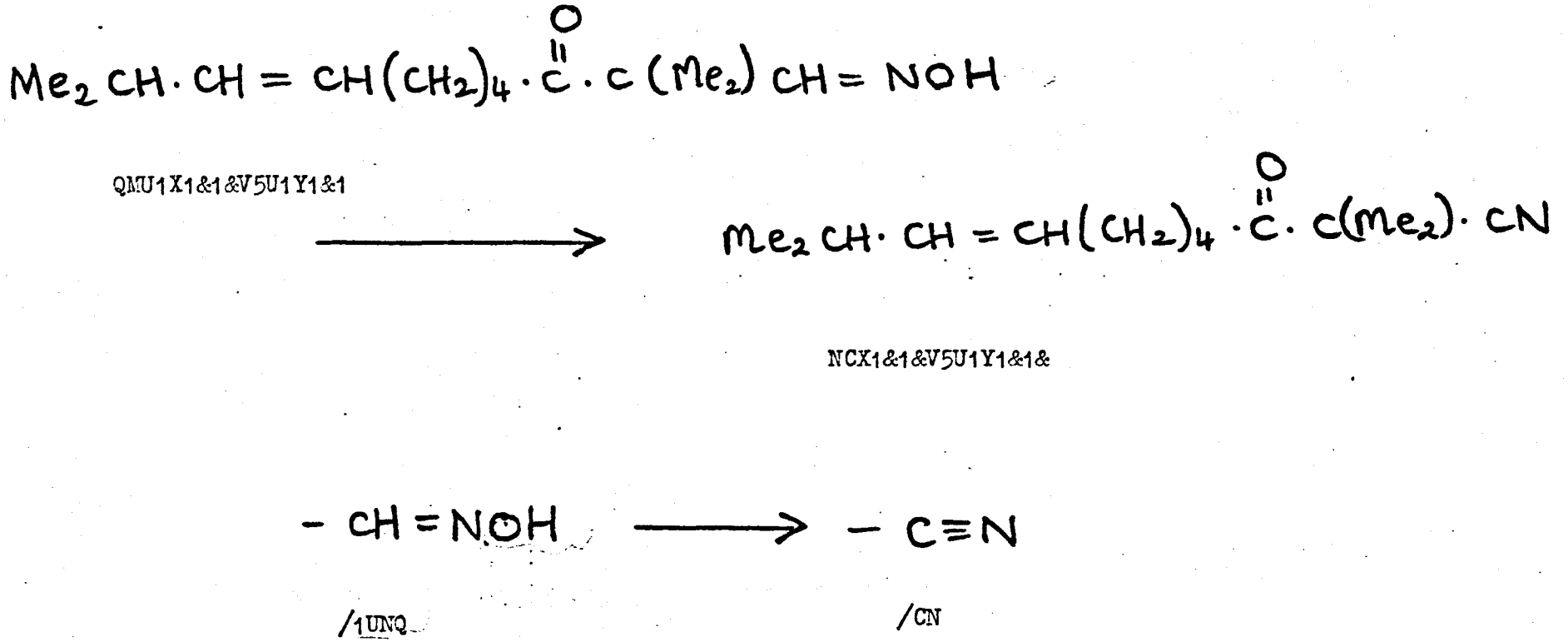


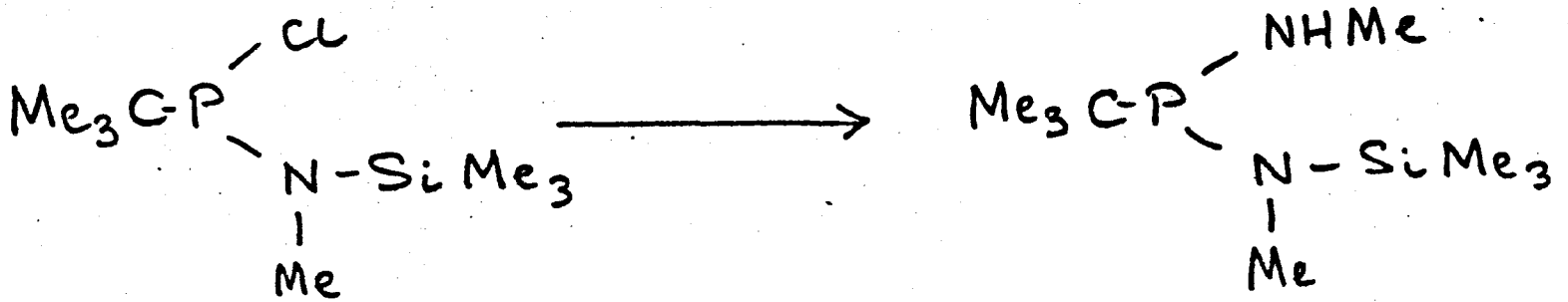
L66 A BTJ



L66TJ

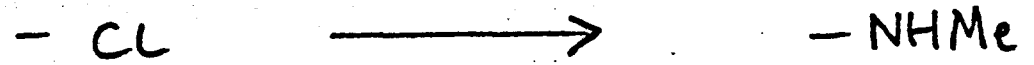



Me2 CH · CH = CH(CH2)4 · C(=O) · C(Me2) CH = NOH
QMU1X1&1&V5U1Y1&1
Me2 CH · CH = CH(CH2)4 · C(=O) · C(Me2) · CN
NCX1&1&V5U1Y1&1&
- CH = NOH
/1UNQ
- C ≡ N
/CN



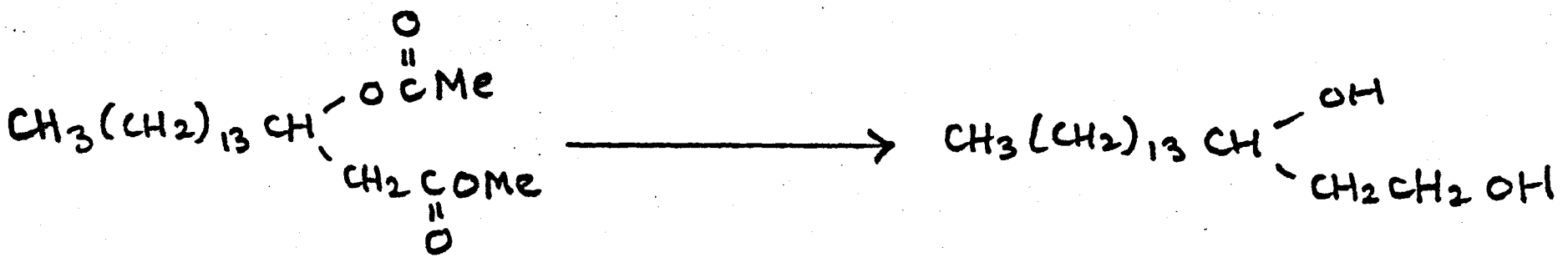
1X1&1&PGN1&-SI-1&1&1

1X1&1&PM1&N1&-SI-1&1&1



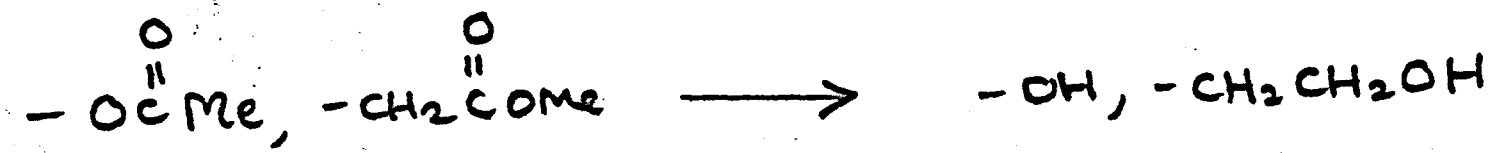
/G

/M1



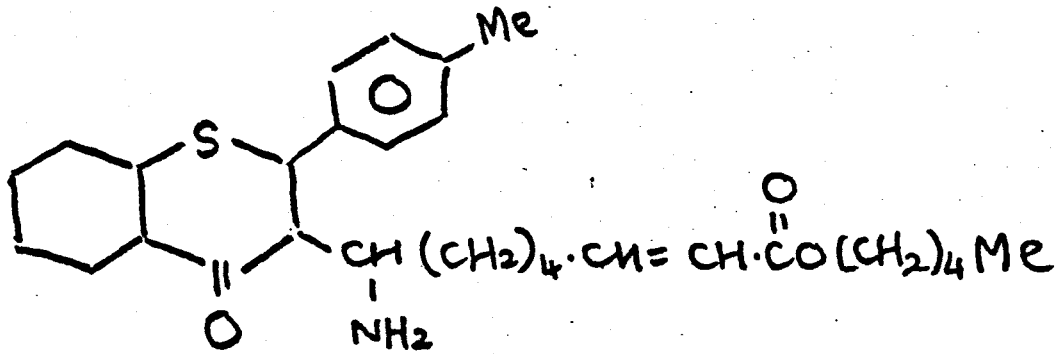
14Y0V1&1V01

Q2YQ14

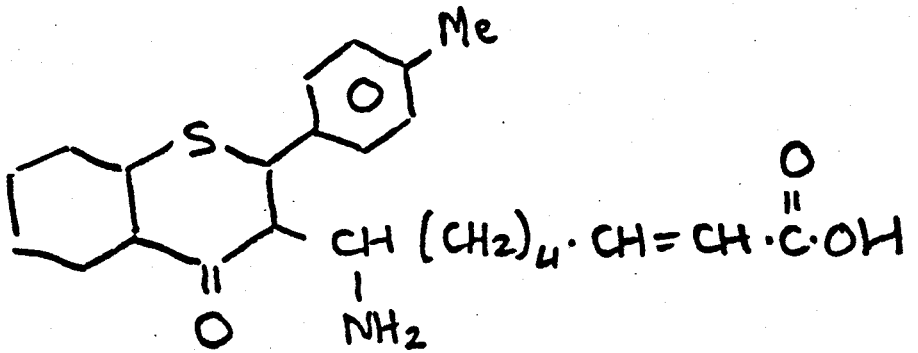


/0V1, /1V01

/Q, /2Q



T66 BS EVTJ CR D1& DYZ5U1V05



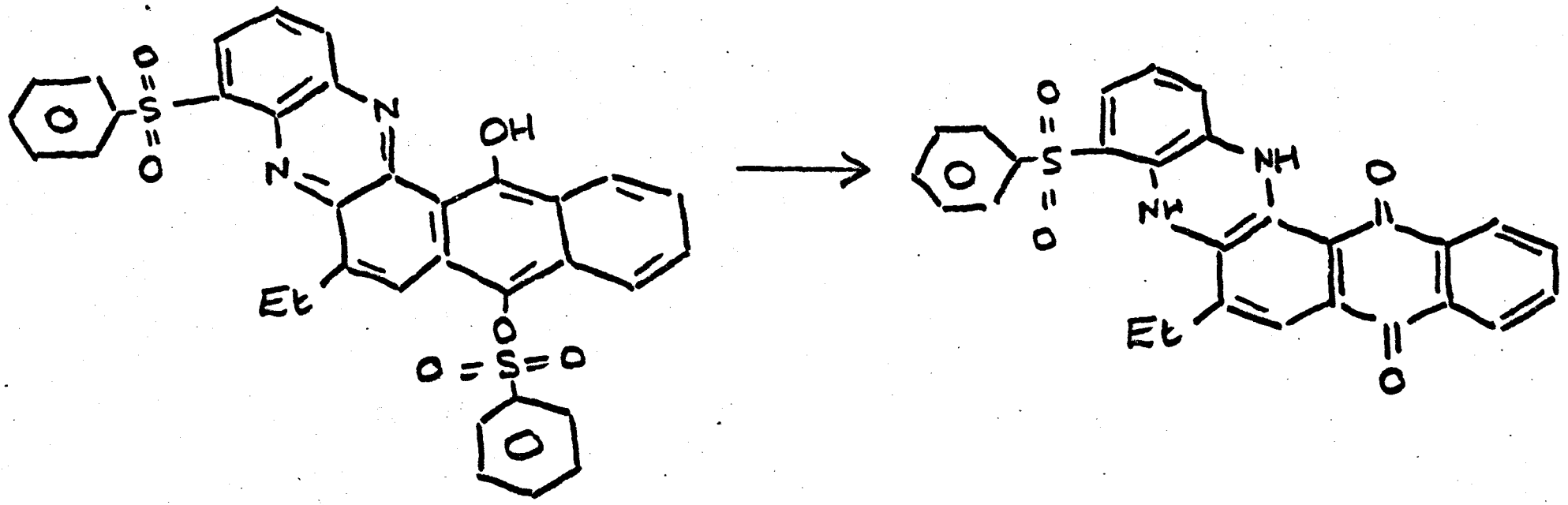
T66 BS EVTJ CR D1& DYZ5U1VQ

/5U1V05

/5U1VQ

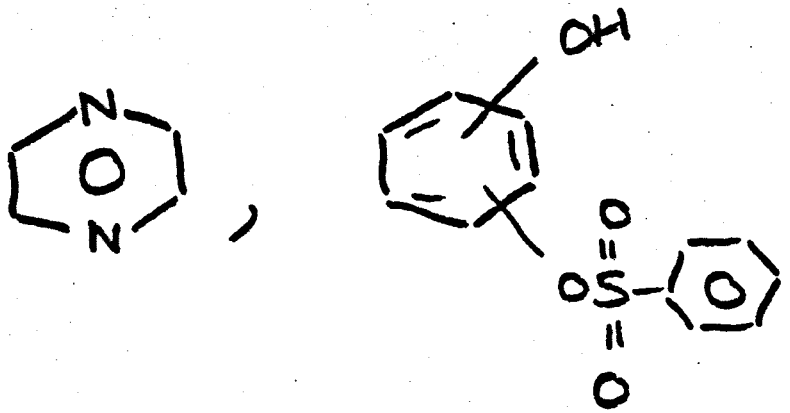
/1U1V01/

/1U1VQ

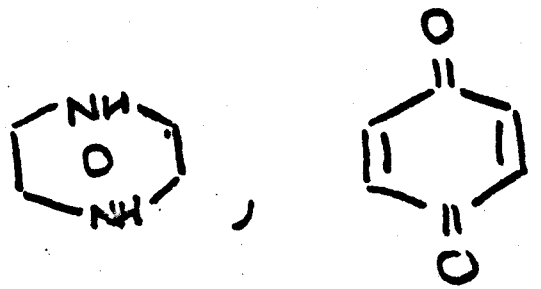


T F6 D6 C666 EN QNJ EQ LOSWR& 02 SSWR

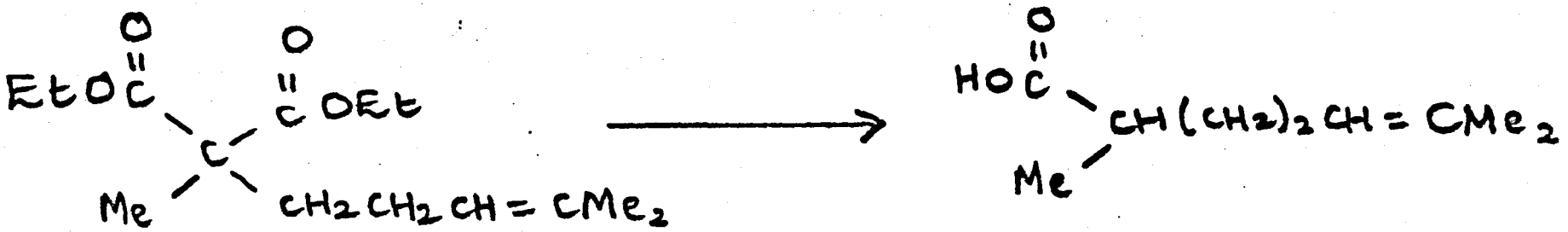
T F6 D6 C666 EM EV LV QMJ 02 SSWR



@T6 AN DNJ, @L6J *Q *OSW* R

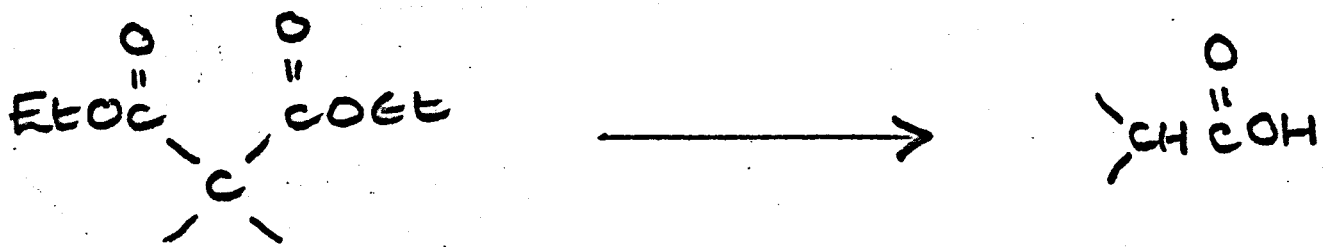


@T6 AM DMJ, @L6 AV DVJ



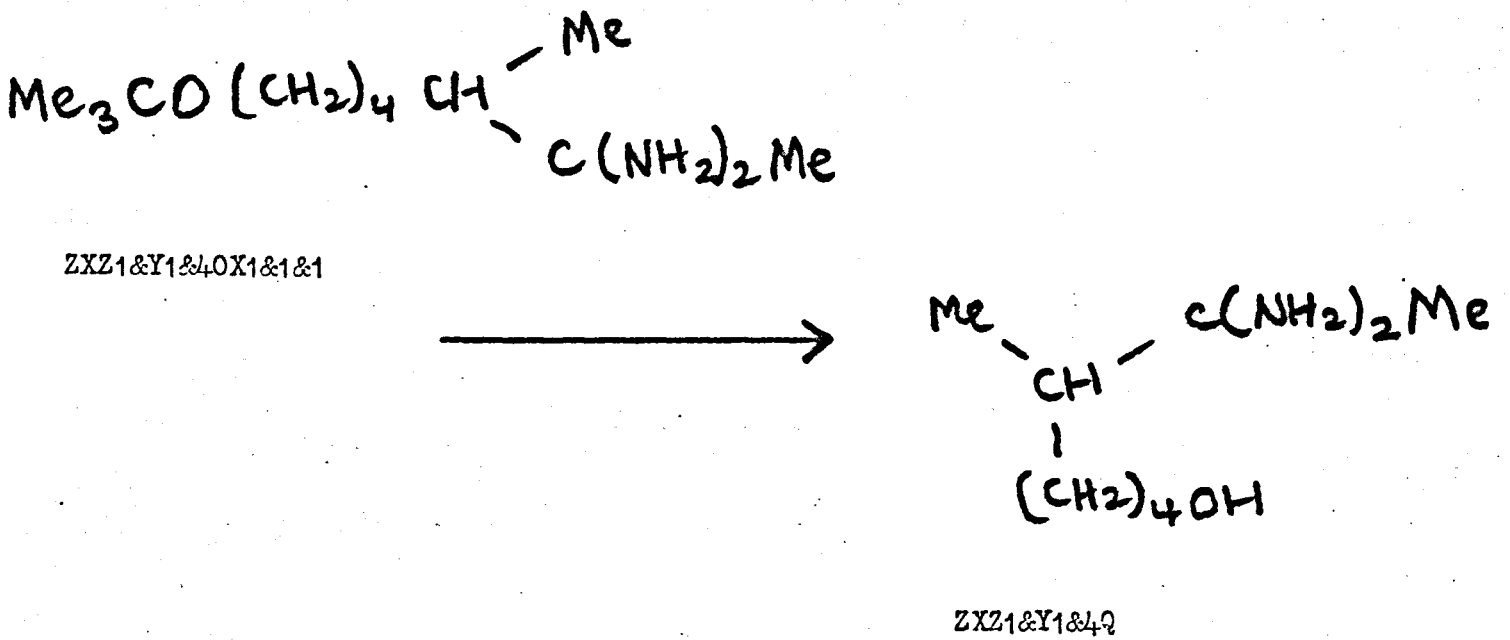
20VX1&V02&3UY1&1

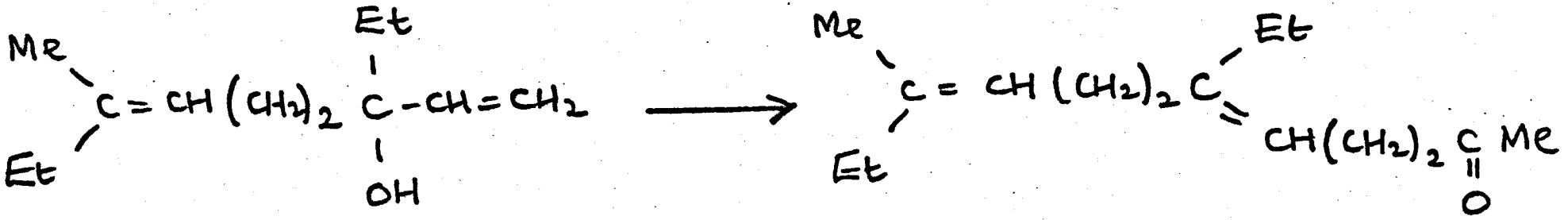
QVY1&3UY1&1



20V/ X /V02

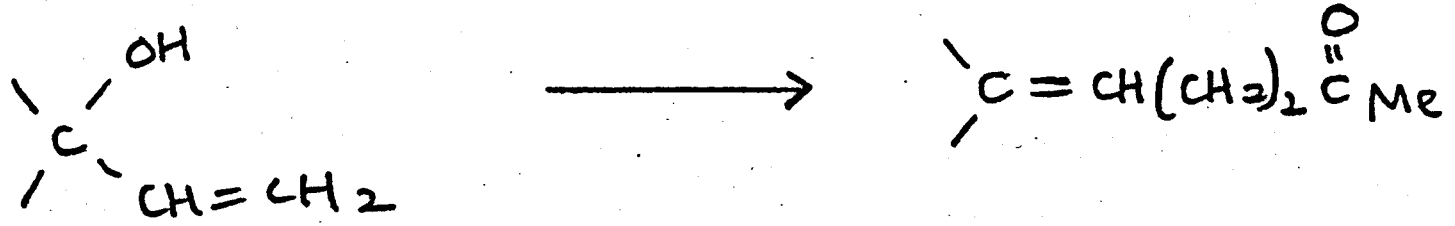
QV/ Y


Me3CO(CH2)4CH
Me
C(NH2)2Me
ZXZ1&Y1&40X1&1&1
Me
CH
C(NH2)2Me
(CH2)4OH
ZXZ1&Y1&40



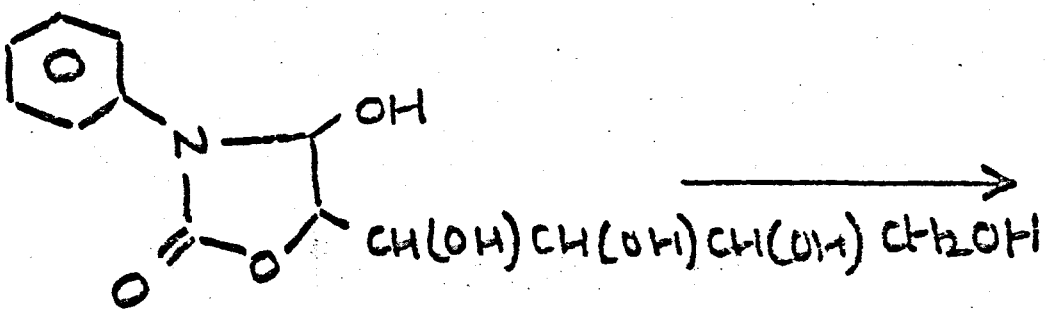
QX2&1U1&3UY1&2

2Y1&U3Y2&U3V1

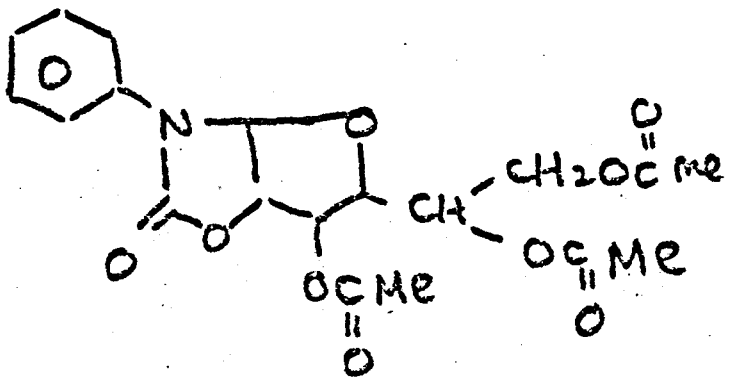


Q/ X /1U1

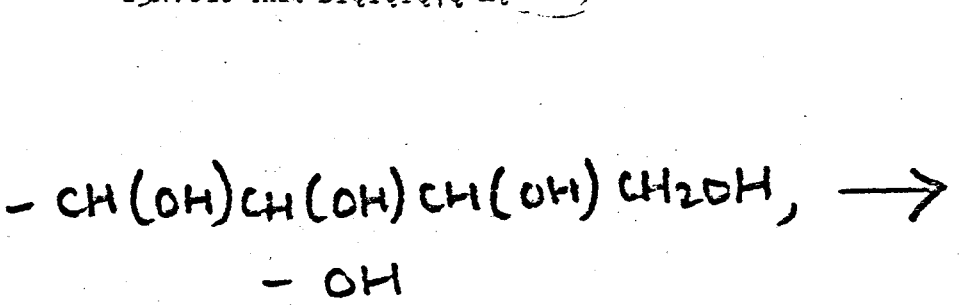
YU3V1



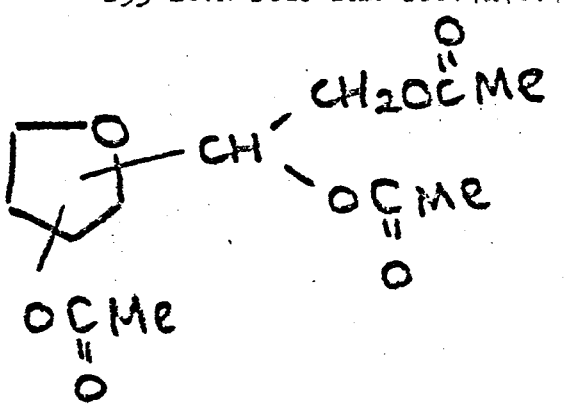
T5NVOTJ AR& DYQYQYQ1Q EQ



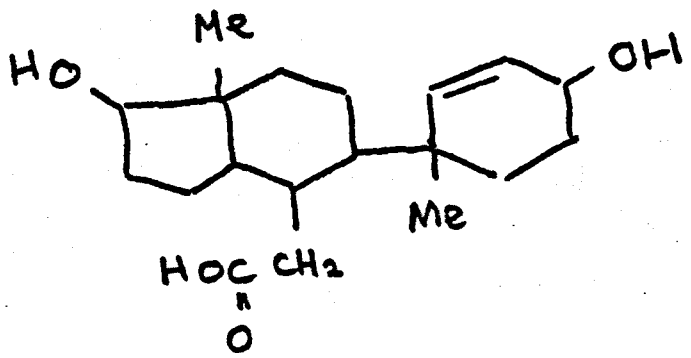
T55 BCVN FOTJ DR& GYCV1&1CV1 HCV1

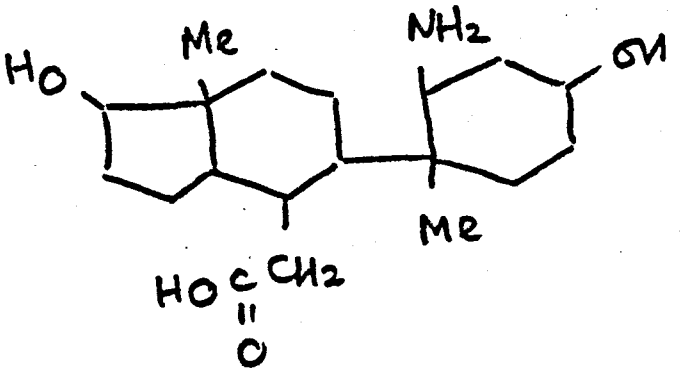


*YQYQYQ1Q, *Q



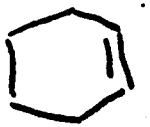
@T5 AOTJ *YCV1&1CV1 *CV1



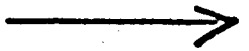


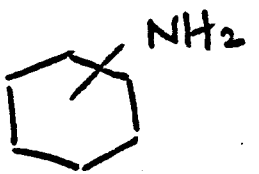
L56TJ A1 BQ F1VQ G- CL6UTJ C1 FQ

L56TJ A1 BQ F1VQ G- AL6UTJ A1 BZ DQ

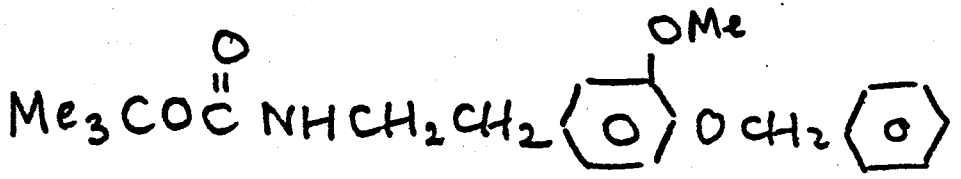


L6UTJ

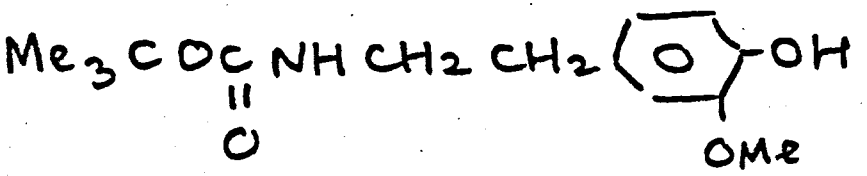




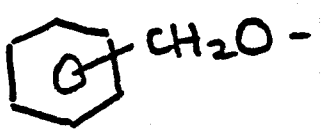
L6TJ *Z



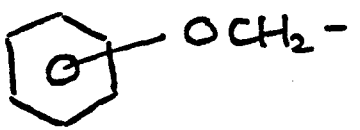
1X1&1&OVM2R CO1 DO1R



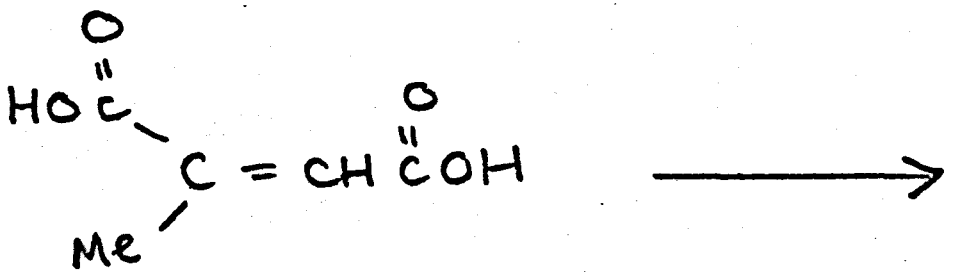
1X1&1&OVM2R DQ CO1



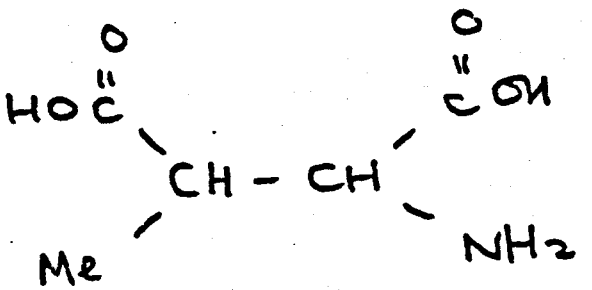
(a) *01* R



(b) R *01*

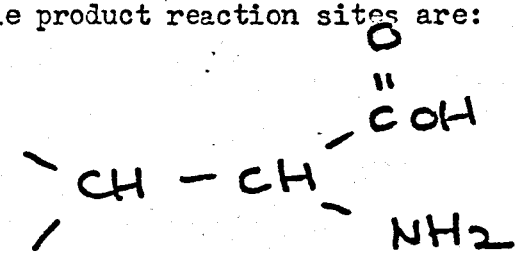


QVY1&1VQ



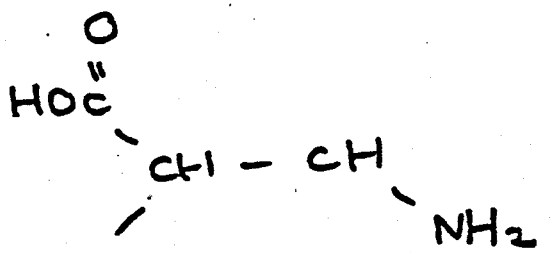
QVYZ1&VQ

The two possible product reaction sites are:

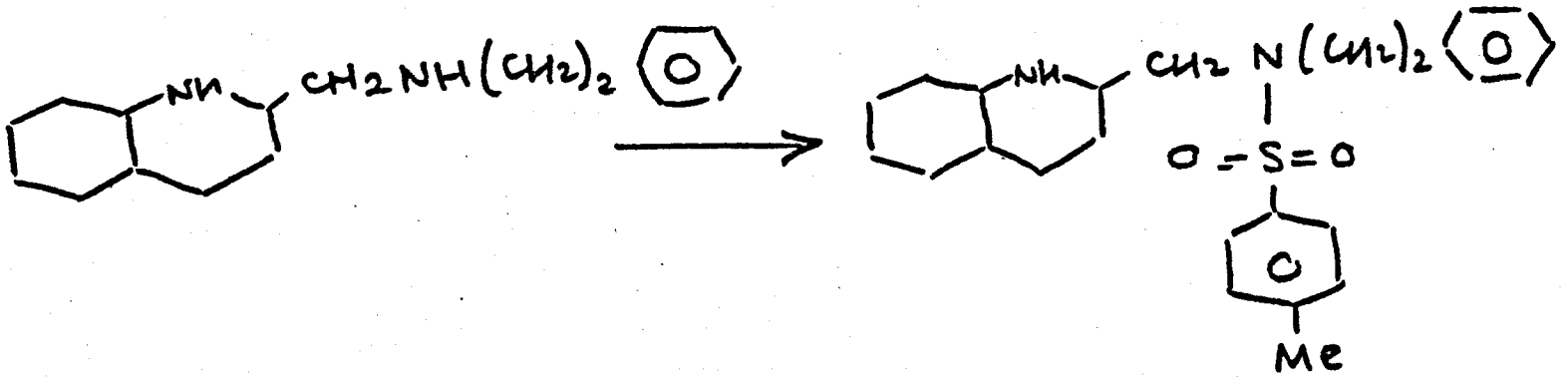


QV/ Y /Z Y

and



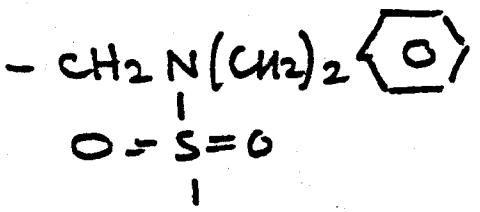
Y /Z Y /VQ



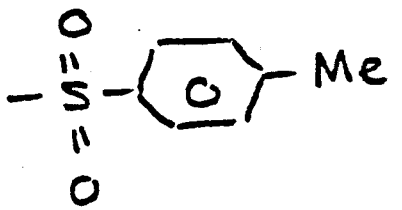
T66 BMTJ C1M2R

T66 BMTJ C1N2R&SWR D1

Since no look-ahead facility is provided, the two possible product reaction sites are

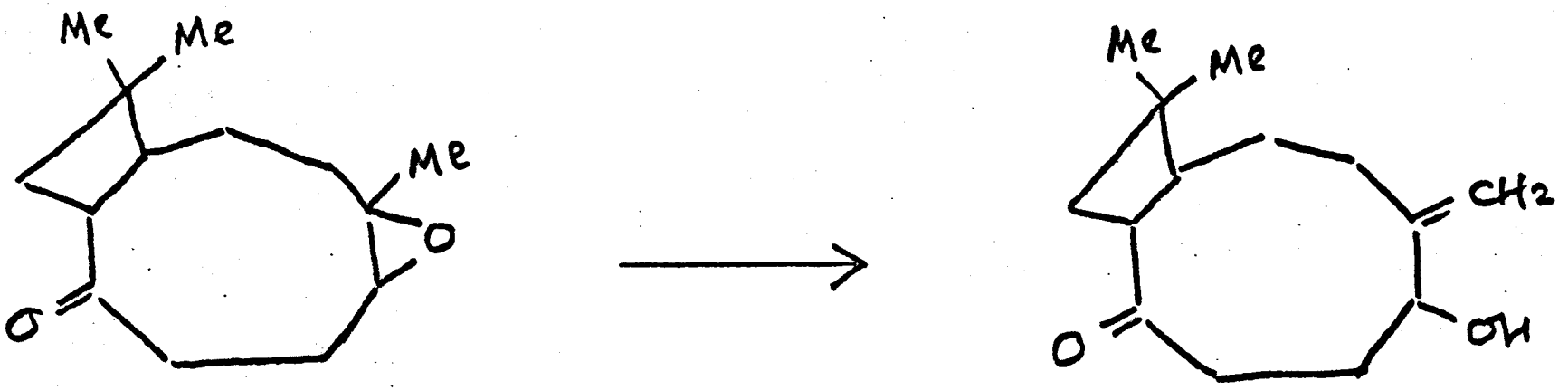


and



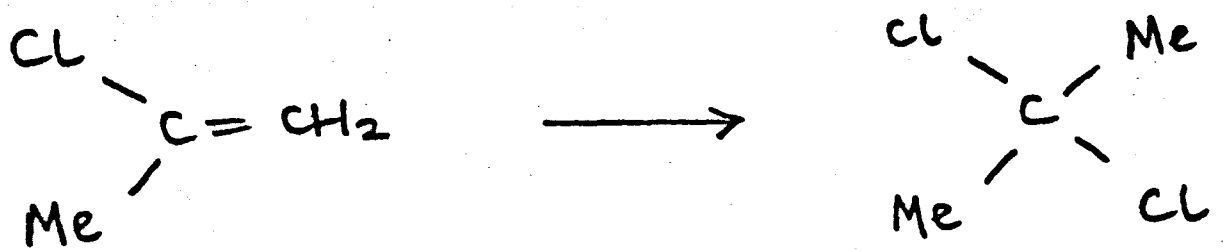
1N2 R /SW*

/SW* R *1



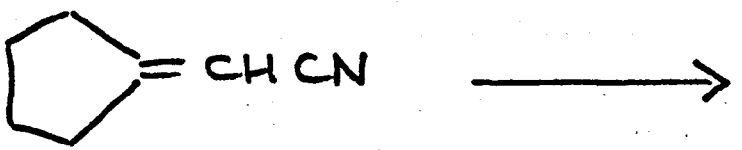
T D394 EO IVTJ D1 L1 L1

L49 EV IYTJ B1 B1 HQ IU1

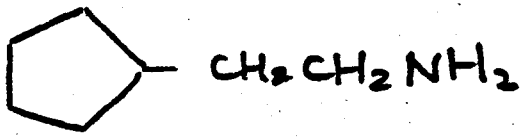


GY1&U1

GX1&1&G

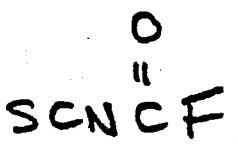


L5TYJ AU1 CN

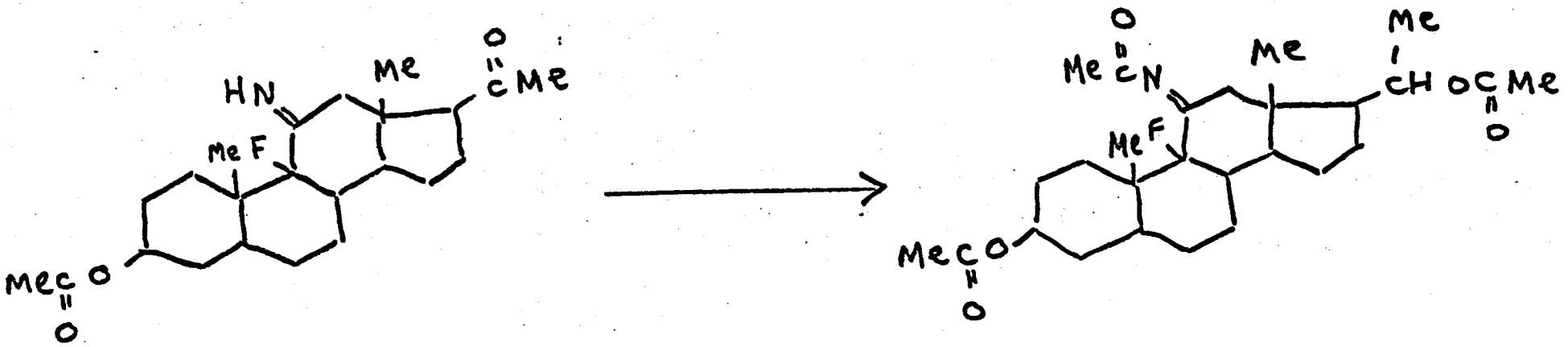


L5TJ A2Z

NCSVF

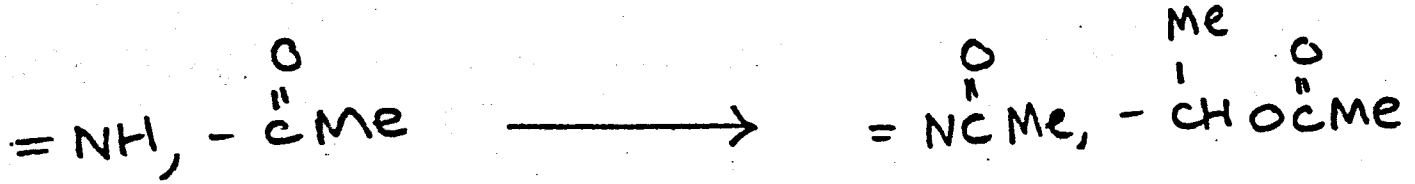


SCNVF



L E5 B666 CYTJ A1 BF CUM E1 FV1 OOV1

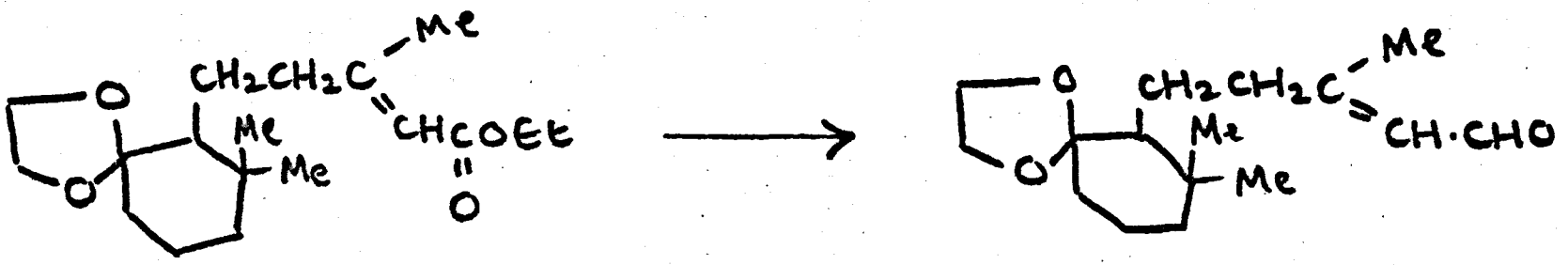
L E5 B666 CYTJ A1 BF CUNV1 E1 FY1&OV1 OOV1



*UM, *V1

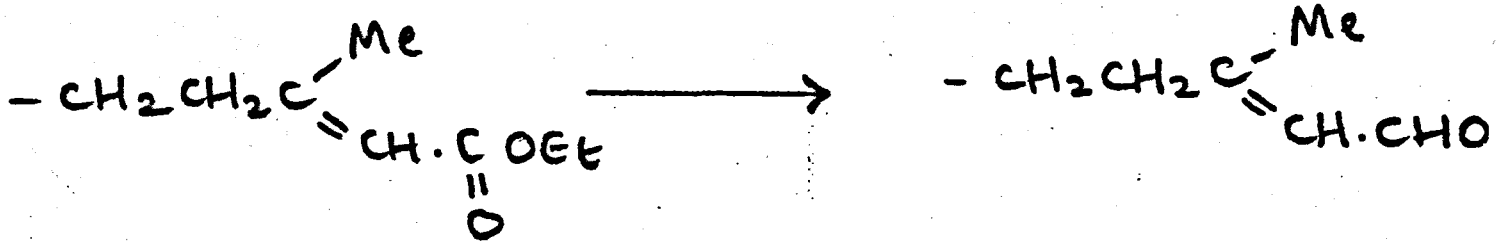
*UNV1, *Y1&OV1

The reactant and product analysis fragments are *UM, *V1 and *UNV1, Y, /1, /OV1 respectively.



T50X0TJ B-& AL6XTJ B2Y1&U1V02 C1 C1

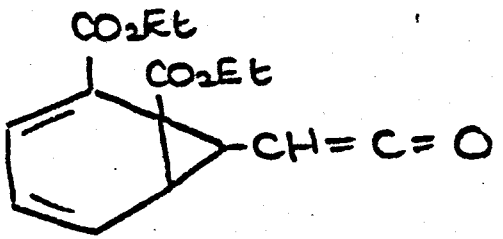
T50X0TJ B-& AL6XTJ B2Y1&U1VH C1 C1

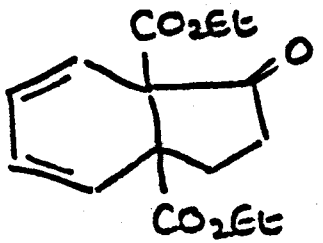


*2Y1&U1V02

*2Y1&U1VH

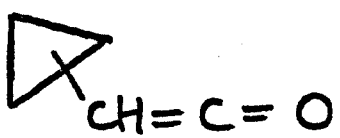
The reactant and product analysis fragments are YU1V02 and YU1VH respectively.

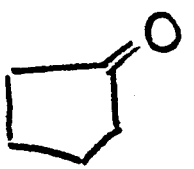




L36 DU FUTJ AV02 B1UC0 GV02

L56 BV CU FU HUTJ AV02 EV02

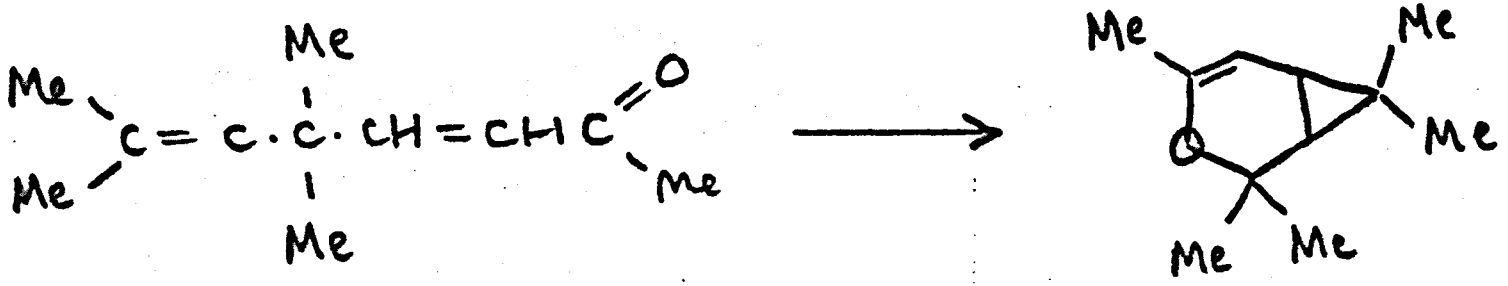




@L3TJ *1UC0

@L5 AVUTJ

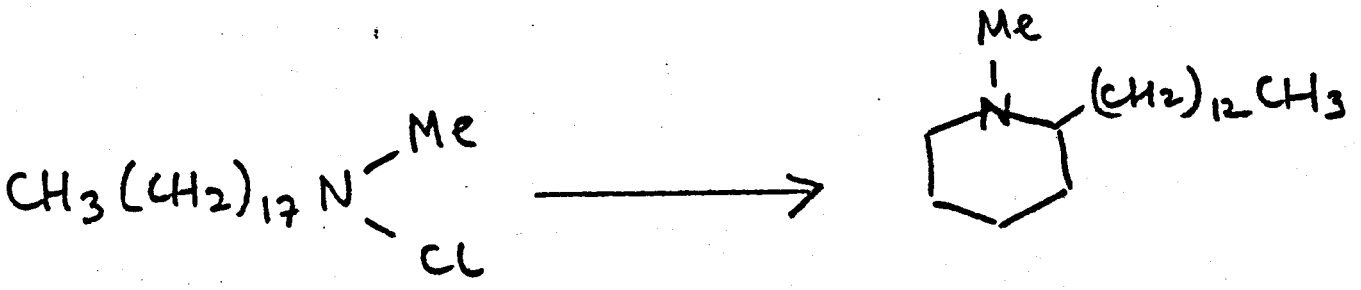
The analysis fragments are those included in the two reaction site notation strings.



1Y1&U1X1&1&1U1V1

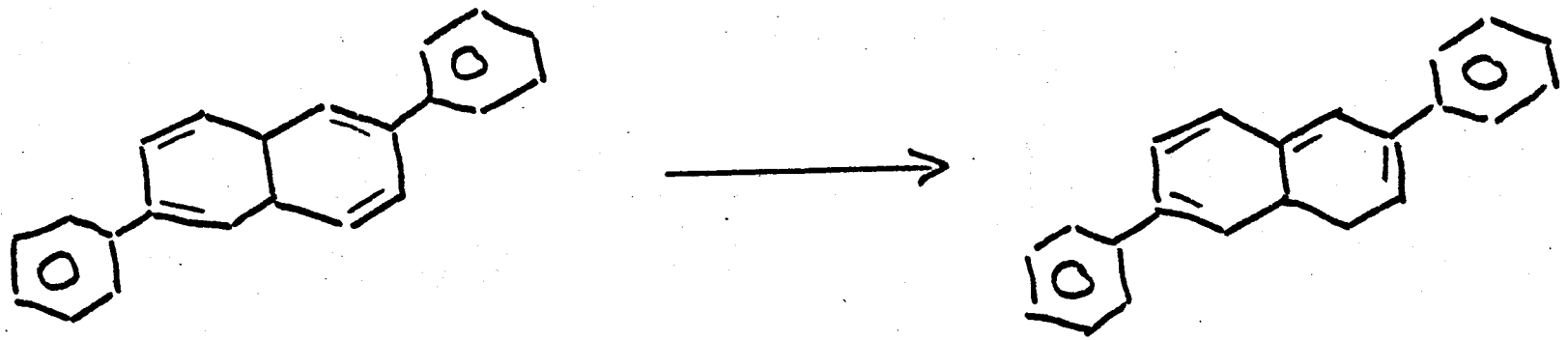
T36 E0 FUTJ B1 B1 D1 D1 F1

The reaction site notations contain all the fragments in the initial structures and the only fragments not common to the two sides of the equation, and which are eliminated in the final analysis, are the five methyl groups since the program identifies as common strings which are identical bar an initial * or / character.



GN1&18 T6NTJ A1 B13

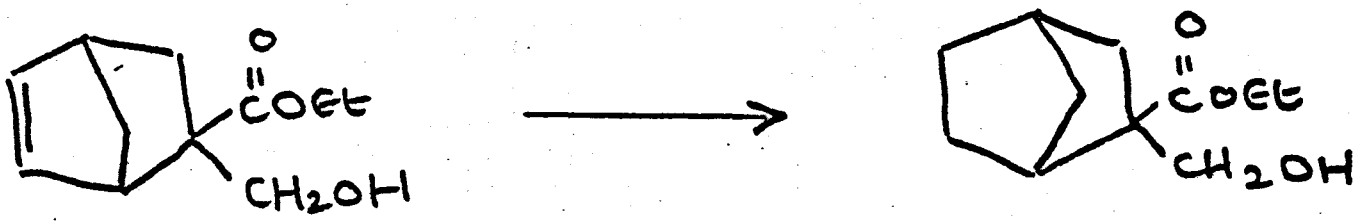
Although the reaction sites would provide little or no useful information, the product analysis fragment T6 ANTJ is probably a sufficient description of the change that has occurred.



L66 BU DU GU IUTJ CR& HR

L66 AU CU FHT&J CR& HR

The reactant and product reaction site notations are @L6UTJ, @L6UTJ and @L6HUTJ, @L6HUTJ since no attempt is made to specifically localise ring saturation features. The analysis fragments would be the same as the members of the reaction sites.



L55 A CUTJ FVO2 F1Q

L55 ATJ CVO2 C1Q

The reaction site notation string is simply @L5UTJ ⟶ @L5TJ.

ORIGINAL COPY TIGHTLY BOUND

121336 2A 05 08↑*

```
*NW            /20SW1   ----->>  *NW20SW1
 P 1           WS1&02NR CG20SW1   ----->>  WS1&02NR CG DNW20SW1
                      122218  4A  03  15↑*

*NW            Z*   ----->>  WN*
 P 1           ZR CSW1G   ----->>  WNR CSW1G
                      122394  1A  02  01↑*

*NW            GVYGUNM* >*E >*E   ----->>  WN* >*1 >*MNUYGVNCS
 P 1           GVYGUNMR BE EE   ----->>  WNR C1 DMNUYGVNCS
                      122632  3A  16  22↑*

*NW            GVYGUNM* >*G >*G   ----->>  WN* >*01 >*MNUYGVNCS
 P 1           GVYGUNMR CG DG   ----->>  WNR C01 DMNUYGVNCS
                      122632  4A  17  23↑*

*NW                 ----->>  *NW
 P 1           T6NNVJ BR DNW& DG FG   ----->>  T6NNVJ BR BNW DNW& DG FG
                      130557  5A  04  06↑*

*NW                 ----->>  *NW
 P 1           T6NNVJ BR DNW& DG EG FG   ----->>  T6NNVJ BR BNW DNW& DG EG FG
                      130557  6A  07  09↑*

*NW                 ----->>  *NW
 P 1           T6NNVJ BR DNW& FG   ----->>  T6NNVJ BR BNW DNW& FG
                      130557  7A  02  03↑*

*NW                 ----->>  *NW
 P 1           T6NNVJ BR DNW& EG FG   ----->>  T6NNVJ BR BNW DNW& EG FG
                      130557  9A  12  14↑*

*NW                 ----->>  *NW
 P 1           T6NNVJ BR DNW& F01   ----->>  T6NNVJ BR BNW DNW& F01
                      130557  :A  15  16↑*

*NW                 ----->>  *NW
 P 1           T6NNVJ BR DNW& D01 FG   ----->>  T6NNVJ BR BNW DNW& D01 FG
                      130557  ;A  17  18↑*

*NW                 ----->>  *NW
 P 1           T6NNVJ BR& DS0&1 FG   ----->>  T6NNVJ BR DNW& DS0&1 FG
                      130561  5A  33  40↑*

*NW            *S0&1   ----->>  *NW ...
```

II.33

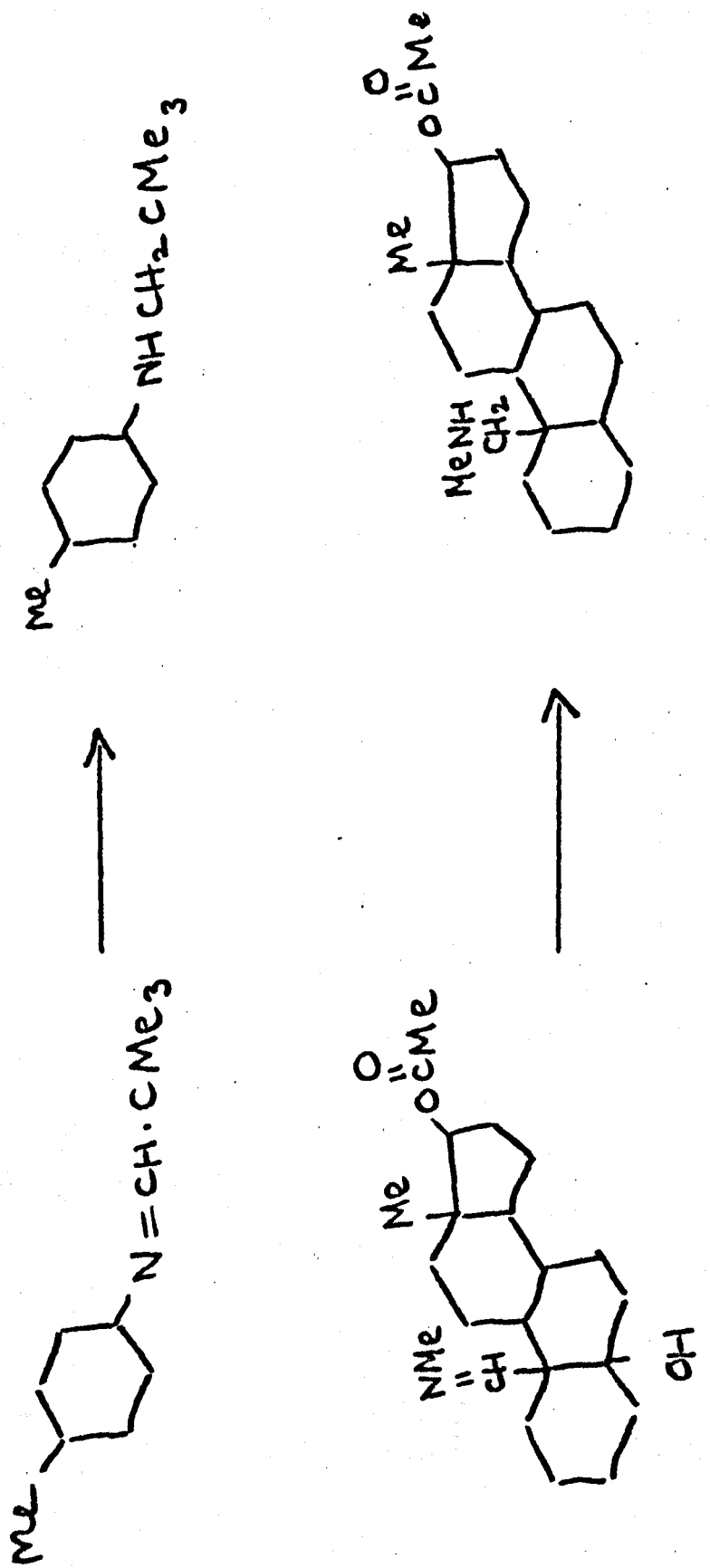
```
R 1                LATJ AD AAVQ A1    ------>> LATJ AD AAVQ1 A1
                   125031  2A  06  08↑*

/1VQ               *YE1VQ    ------>>  *1U1VQ
  R 1              T56 BNSNJ FYE1VQ    ------>>  T56 BNSNJ F1U1VQ
                   125393  2A  01  06↑*

/1VQ               *YE1VQ    ------>>  *1U1
  R 1              T56 BNSNJ FYE1VQ    ------>>  T56 BNSNJ F1U1
                   125393  3A  01  06↑*

/1VQ               *S2VQ     ------>>  #T6 AS DVTJ
  R 1              T56 BSJ DS2VQ       ------>>  T B666 CS FV HST&RJ
                   125618  2A  06  07↑*

/1VQ               QV1Y* /1HV*     ------>>  T5 AN BVTJ *V*
  R 1              QV1YR D01&1HVR D02      ------>>  T5HVTJ AVR D02& DR D01
                   125628  4A  32  02↑*

/1VQ               QV1Y* /1HV*     ------>>  T5 AN BVTJ *V*
  R 1              QV1YR D02&1HVR C1 D01   ------>>  T5NVTJ AVR C1 D01& DR D02
                   125628  5A  36  06↑*

/1VQ               QV1/ X /1 /Q /2Q    ------>>  #L6J *1 >#T6 AN DVJ *1 >#L6J
  R 1              QV1X1&Q2Q &&3/C-14      ------>>  T C666 BN IVJ B1 E1 &&1&/C-14
                   129286  1A  06  07↑*

/1VQ               QV2*      ------>>  #L5 AVTJ
  R 1              QV2R CQ FE D01      ------>>  L5A BVT&J FE H01 IQ
                   129397  3A  11  14↑*

/1VQ               QV3HV1*     ------>>  T5 AN BVTJ *YH1* >/N1&1
  R 1              QV3HV1R     ------>>  T5NVTJ AYU1R&N1&1
                   129404  6A  34  32↑*

/1VQ               *3VQ      ------>>  #L6 AVTJ
  R 1              L66J C3VQ H2VQ      ------>>  L B666 CVT&&J L2VQ
                   129444  4A  03  06↑*

/1VQ               *2VQ      ------>>  #L5 AVTJ
  R 1              L B666&&TJ E2VQ     ------>>  L E5 B666 HVT&&TJ
                   129444  8A  08  07↑*

/1VQ               *Y1* /1VQ       ------>>  #L6 AVTJ
  R 1              T56 BVNVJ CY1R&1VQ      ------>>  T5A BVNVJ C- DL66 BVT&J
                   130313  3A  03  04↑*
```

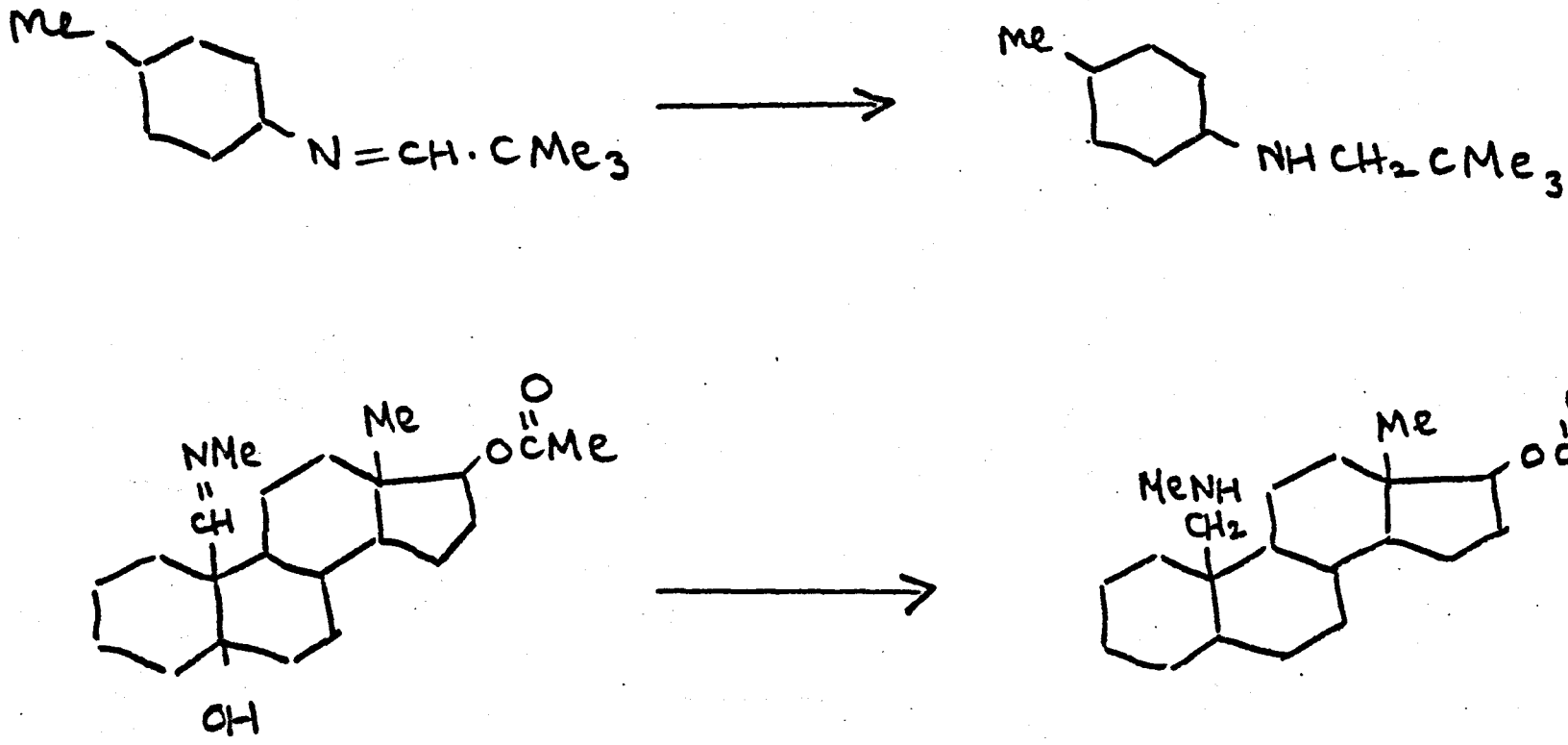
II.33 cont.

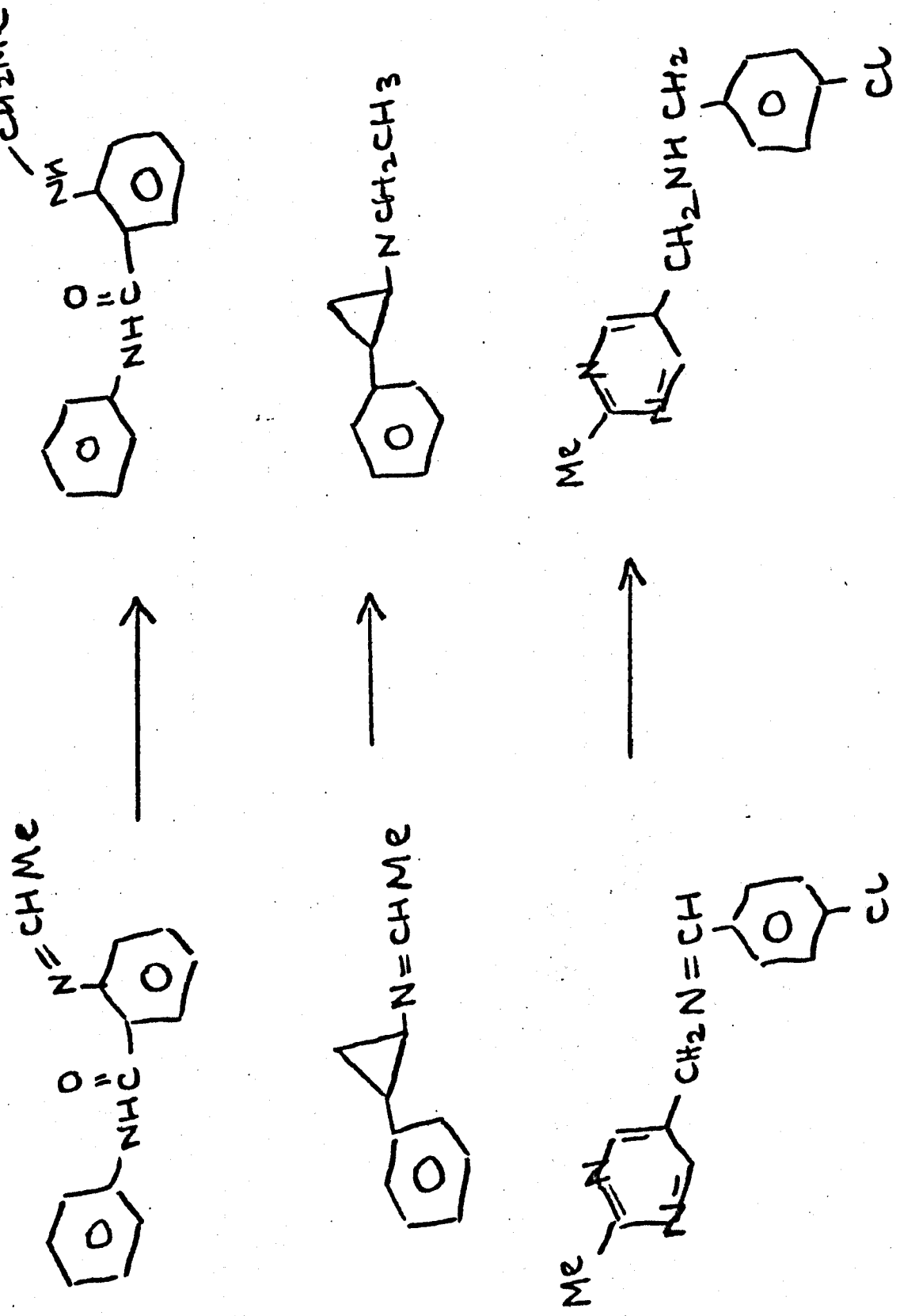
```
P 1                     13 AN BV CNTJ 0E 0F
                        T5NVNJ A1R& CR& DR& EE      ----->>  T5NVNTJ A1R& CR& DE DR& EE FF
                                  121427  <A  15  20↑*

T5 AN BV CNTJ           T6 AN BV CN DVJ      ----->>  T5 AN BV CNTJ *VQ
  P 1                   T6NVNVJ A1 C1R& EQ      ----->>  T5NVNTJ A1 C1R& DVQ DQ
                                  123380  3A  06  07↑*

T5 AN BV COTJ           T5 AN BV COTJ *YQYQYQ1Q        ----->>  T6 AOTJ *OVH* *OV1 *OV1 *1OV1
  R 1                   T5NVOTJ AR& DYQYQYQ1Q EQ    ----->>  T6OTJ BQ COVHR& DOV1 EOV1 F1OV1
                                  123384  2A  01  07↑*

T5 AN BV COTJ           VH1OVH*      ----->>  T5 AN BV COTJ *Q
  P 1                   VH1OVHR      ----->>  T5NVOTJ AR& EQ
                                  123384  ;A  19  11↑*

T5 AN BV COTJ           VH1OVH*      ----->>  T5 AN BV COTJ *OV1
  P 1                   VH1OVHR      ----->>  T5NVOTJ AR& EOV1
                                  123384  ;A  19  12↑*

T5 AN BV CSJ            T5 AN BV CSJ *MYUS&* >*VQ     ----->>  T8 AS BV CS EN FNHJ
  R 1                   T5NVSJ AMYUS&R DVQ& E1      ----->>  T8SVS ENN HHJ DR& G1
                                  124177  3A  03  04↑*

T5 AN BV CSJ            #T5 AK BN DSJ >#T5 AK CSJ >*VO1     ----->>  T5 AN BV CSJ *MYUS&* >*VQ
  P 1                   T55 AKN DS CSJ CR DVO1& H1 && H-S-04    ----->>  T5NVSJ AMYUS&R DVQ& E1
                                  124177  2A  02  03↑*

T5 AN BV CY DSJ         T5 AN BV CY DS EYJ *US        ----->>  T5 AN BV CY DSJ *S1
  P 1                   T5NVYSYJ CU1N1&1 EUS      ----->>  T5NVYSJ CU1N1&1 ES1
                                  125630  1A  08  09↑*

T5 AN BV CYTJ           T6 AO DYJ *U* >T5 AN BV CYTJ *1        ----->>  T6 AOJ *G
  R 1                   T6O DYJ BR& FR& DU- CT5NVYTJ A1     ----->>  T6OJ BR& DG FR && P-02-G2 &3/20
                                  134063  3A  07  17↑*

T5 AN BV DMHJ           T5 AN BV DMTJ *Z        ----->>  T5 AN BV DMHJ
  P 1                   T5NV DMTJ EZ E- DT5M CNJ ENU1Q     ----->>  T5NV DM CHJ E- DT5M CNJ ENU1Q
                                  127139  3A  12  13↑*

T5 AN BV DSHJ           YUM >/S1VQ      ----->>  T5 AN BV DSHJ *YUM&S1VQ
  P 1                   QV1SYUM&YUM&S1VQ      ----->>  T5NV DS CHJ EYUM&S1VQ
                                  119798  1A  12  01↑*

T5 AN BVHJ              T5 AN BVHJ *V1       ----->>  T5 AN BVHJ
  R 1                   T5NV EHJ AV1 EVO2 EVO2      ----->>  T5NV EHJ EVO2 EVO2
                                  132603  3A  01  05↑*
```

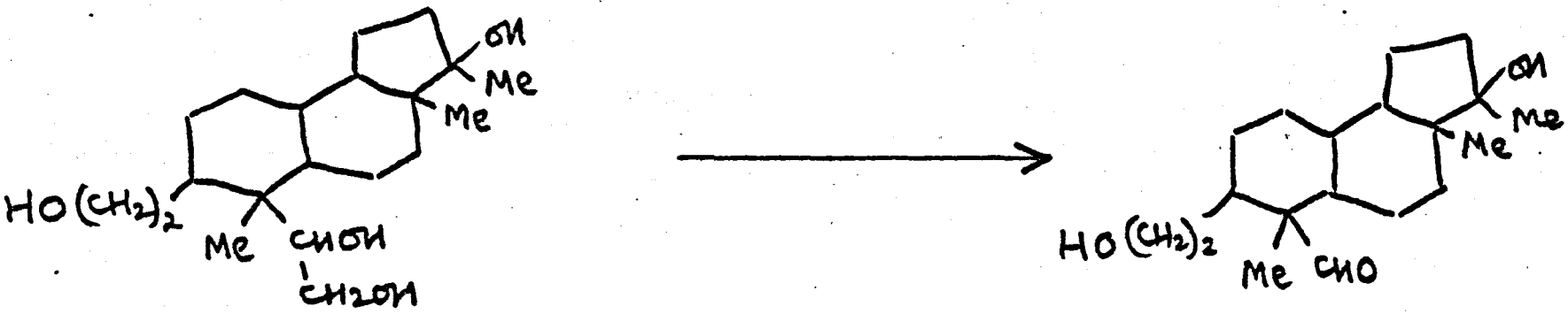
```
#T6 AN OVTJ          #T6 ANHJ *Q *VO1      ------>> #T6 AN OVTJ
  P 1                T C6566 1A P BN LN PHJ NQ OVO1 P1    ------>> T C6566 1A P BN LN NV&&RTJ P1
                              131210  1A  01  02↑*

#T6 AN DVUTJ         #T6 ANUTJ *01 >#L3TJ      ------>> #T6 AN DVUTJ >#L3 AVTJ
  P 1                T636 1A J CN BU EU IUTJ BO1 XO1   ------>> T636 1A J CN FV BU IUTJ BO1
                              122706  9A  07  08↑*

#T6 AN DVUTJ         *X1&02&02 >*1V02     ------>> #T6 AN DVUTJ *02
  P 1                T6NTJ AX1&02&02 B1V02     ------>> T66 AN DV BUTJ BO2
                              131928  7A  05  07↑*

#T6 AN DVUTJ         L8HJ *VNSUG          ------>> #T6 AN DVUTJ >#T8 ANUTJ
  P 1                L8 AHJ AVNSWG BO1 &&2/10     ------>> T68 A B AN DV AU GU IUTJ BO1
                              121287  <A  23  14↑*

#T6 AN DYJ           #T6 AN DYJ *U1        ------>> #T6 AKJ *1
  R 1                T C666 BN IYJ B3H1 IU1     ------>> T C666 BKJ B3M1 I1 &&G
                              121984  1A  11  13↑*

#T6 AN DYJ           #T6 AN DYJ *U1        ------>> #T6 AKJ *1
  R 1                T C666 BN IYJ B3N1&1 EG IU1     ------>> T C666 BKJ B3N1&1 EG I1 &&G
                              121984  2A  12  14↑*

#T6 AN DYJ           #T6 AN DYJ *U1        ------>> #T6 AN DVJ
  R 1                T C666 BN IYJ B3H1 IU1     ------>> T C666 BN IVJ B3M1
                              121984  3A  11  15↑*

#T6 AN DYJ           #T6 AN DYJ *U1        ------>> #T6 AN DVJ
  R 1                T C666 BN IYJ B3N1&1 EG IU1     ------>> T C666 BN IVJ B3N1&1 EG
                              121984  4A  12  16↑*

#T6 AN DYJ           #T6 AN DYJ *U1        ------>> #T6 ANHJ *1
  R 1                T C666 BN IYJ B3H1 IU1     ------>> T C666 BN IHJ B3M1 I1
                              121984  5A  11  17↑*

#T6 AN DYJ           #T6 AN DYJ *U1        ------>> #T6 ANHJ *1
  R 1                T C666 BN IYJ B3N1&1 EG IU1     ------>> T C666 BN IHJ B3N1&1 EG I1
                              121984  6A  12  18↑*

#T6 AN DYJ           #T5 AN CSTJ >#T6 AN DYJ *US       ------>> #T5 AK CSTJ *0 >#T6 AKJ *SWO
  R 1                T56 AN DS GYT&J FQ GUS I1     ------>> T56 AK DST&J DO FQ GSWO I1 &&6/23
                              125179  1A  10  07↑*

#T6 AN DYJ           #T5 AK CSTJ >#T6 AKJ *0 *E        ------>> #T5 AN CSTJ >#T6 AN DYJ *0 *US
  P 1                T56 AK DST&J FO GE I1 &&6/15      ------>> T56 AN DS GYT&J FO GUS I1
```

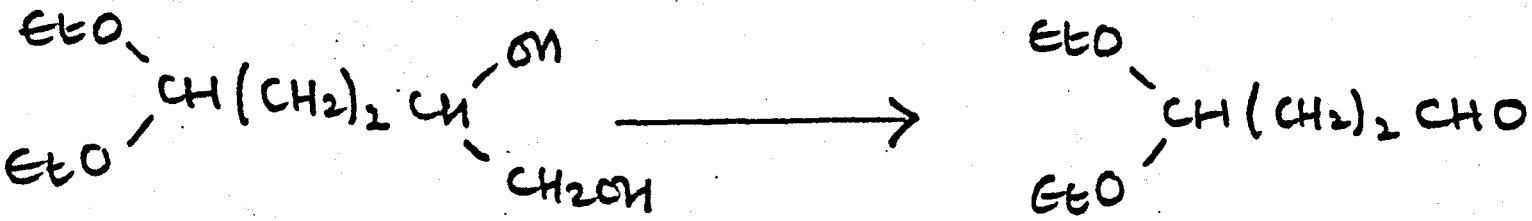
II.33 cont.

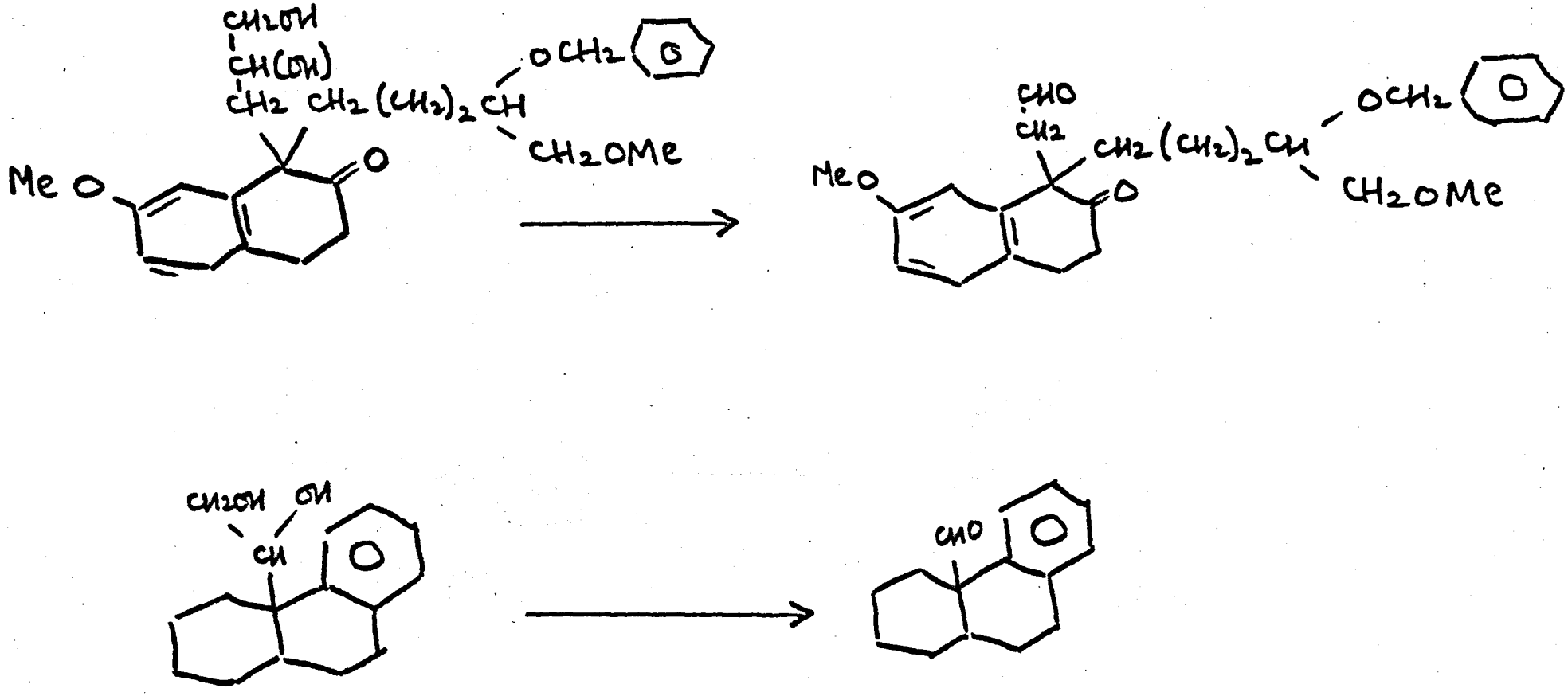

Me
NH CH2 CMe3
Me
NeNH CH2
O
OCMe
N=CH·CMe3
Me
NMe
O
OCMe
Me
OH

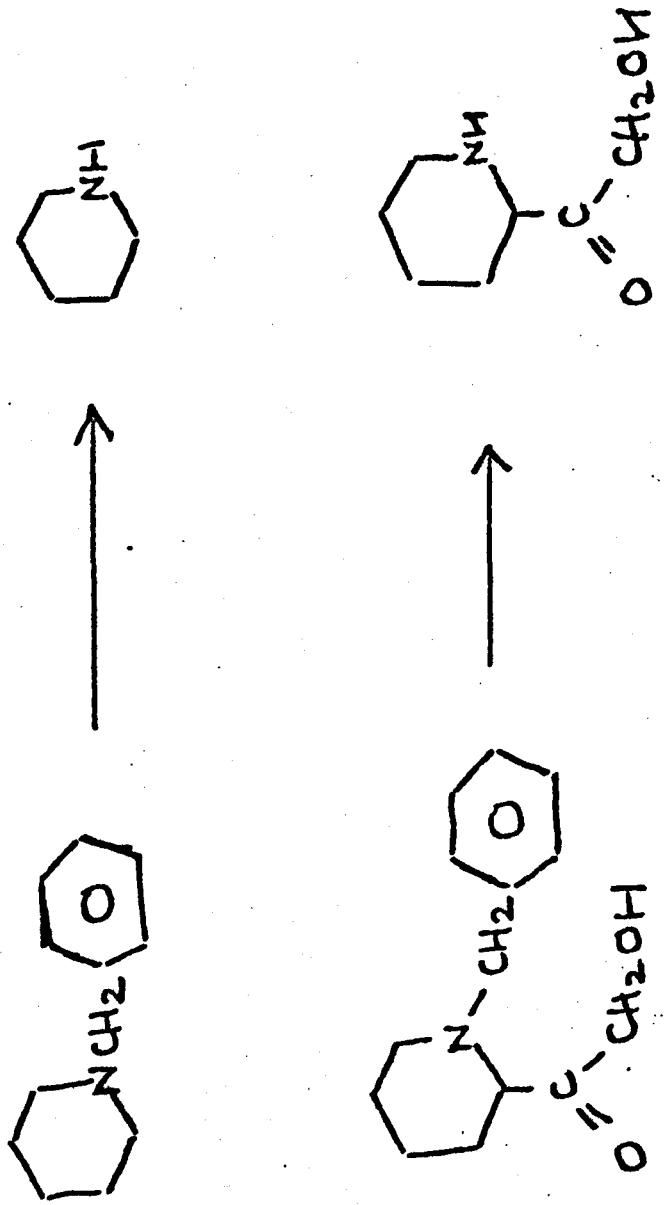

Me
N=CH·CMe3
Me
NH CH2 CMe3
NMe
CH
Me
O
OCMe
OH
MeNH
CH2
Me
O
OCMe

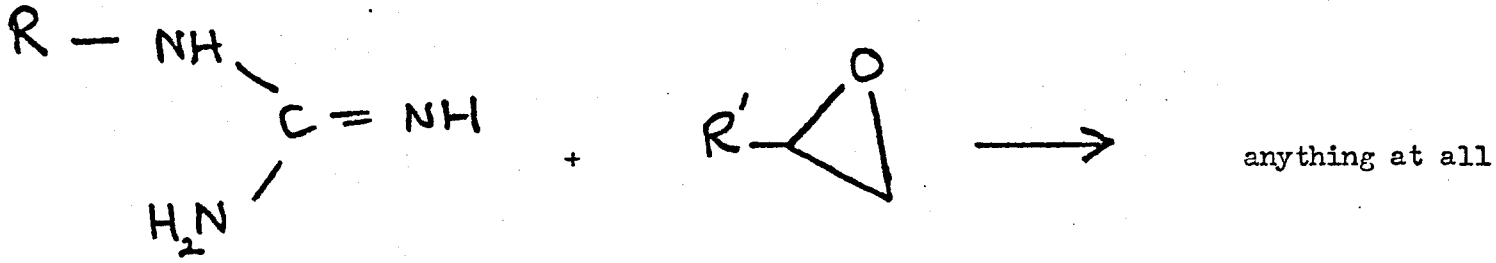

CH2Me
N
O
O=C
NH
CHMe
N=CHMe
N CH2CH3
CH2NHCH2
Cl
Me
CH2N=CH


OH
Me
Me
HO(CH2)2
Me
CHOH
CH2OH
OH
Me
Me
HO(CH2)2
Me
CHO

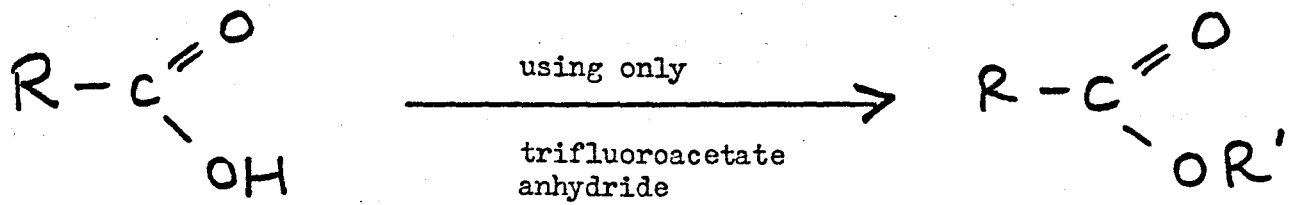

EtO
CH(CH2)2CH
OH
EtO
CH2OH
EtO
CH(CH2)2CHO
EtO

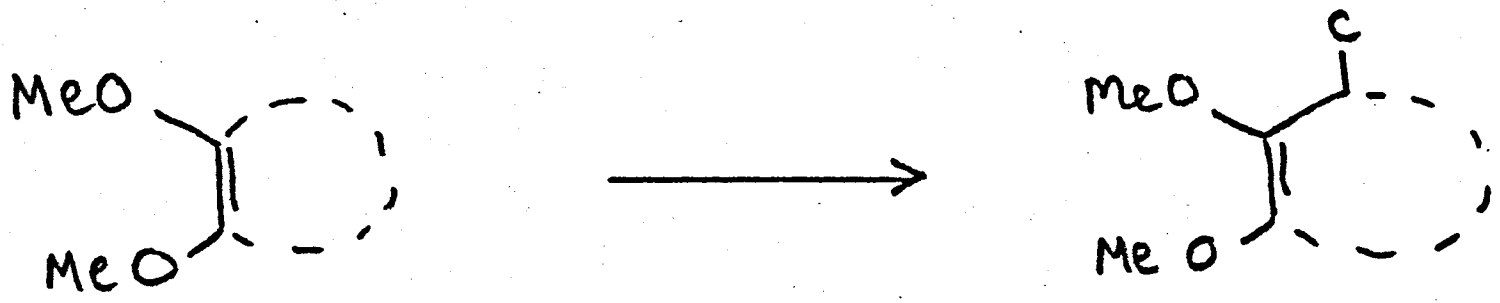

CH2OH
CH(OH)
CH2
CH2(CH2)2CH
OCH2
O
CH2OMe
MeO
O
CHO
CH2
CH2(CH2)2CH
OCH2
O
CH2OMe
MeO
O
CH2OH
OH
CH
O
CHO
O


NH
O
CH2
N
NH
C
O
CH2OH
O
CH2
N
C
O
CH2OH



anything at all


R – C
O
OH
using only
trifluoroacetate
anhydride
R – C
O
O R'



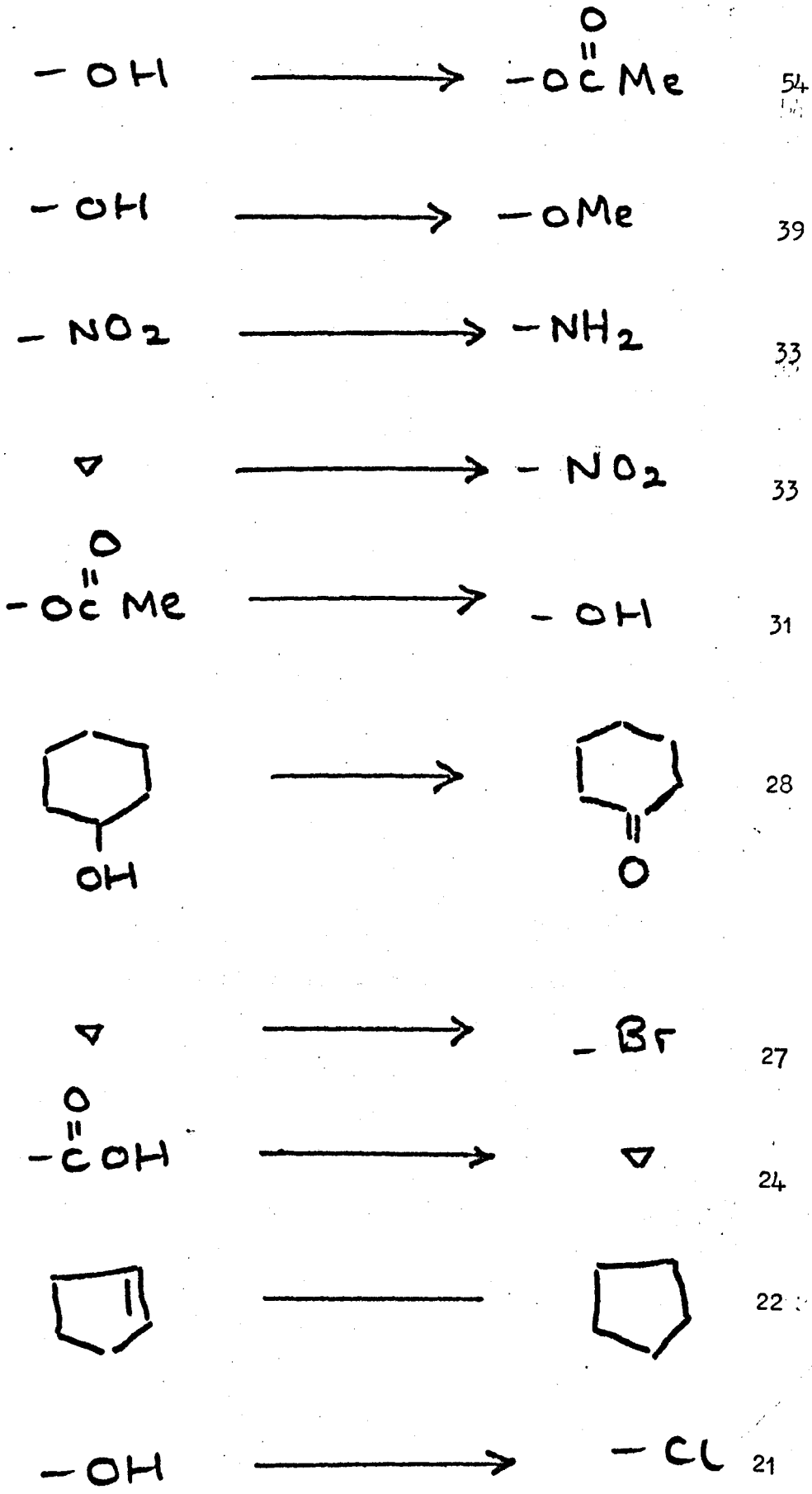
The dotted circle represents any ring (system).

The ten most common analysis fragments.

The numbers after each fragment type correspond to the reactant, product and combined frequencies.

*Q	796	738	1534	*1	282	449	731
/1	486	703	1189	/Q	332	321	653
/1/	415	631	1046	*1/	244	336	580
Y	555	473	1028	*OV1	228	285	513
R	445	486	931	*Z	217	239	456

It will be noticed that many of these fragments arise primarily from the fragmentation methods used and would not actually be used for searching a printed index.



All substituents shown are upon rings and all the rings are fused. The numbers in the right hand column are frequencies of occurrence in the complete file of 7415 reaction site notations.

CHAPTER III

The use of connection table records in the automatic analysis of chemical reaction data.

III.1 An approximate structure matching algorithm

We have earlier described a method for automatic reaction indexing suggested by Vleduts(114) which involves the identification of the maximal subgraphs common to two graphs, i. e., to the connection tables representing the reactant and product molecules of the reaction. It was stated that the algorithm would probably be limited to structures not exceeding ten to fifteen atoms, and Vleduts accordingly described a procedure whereby a comparison of the WLN symbol strings of the reacting molecules would be used to provide 'guiding information', i. e. reactant-product atom equivalences, to reduce the amount of iterative atom matching that would need to be performed. The approximate structure matching algorithm presented in this chapter was initially developed to provide an alternative means of obtaining this guiding information but we now feel that the method can, in solo, be used to process large files of chemical reactions and to provide machine-readable representations for search which could then be interrogated using currently available (sub)structure search techniques. The procedure consists of identifying large areas common to both sides of the reaction equation; taken together, these areas may correspond to the maximal common subgraph(s) but this will not generally be so. Accordingly, it is not possible to delineate specifically the bonds changed by the reaction but this limitation is more than offset by the very much larger number of reactions that may be processed in the same amount of time.

We consider chemical structures as represented by labelled graphs the nodes of which are the atoms and the edges the bonds. The graphs R and P are the reactant(s) and product(s) of a chemical reaction and their nodes are denoted by r_i and p_i, or generally a_i. The simple reaction of Fig. III.1 will be used to illustrate

the basic procedure, the aim being to isolate the change shown in the lower part of the Figure and to note that reactant atoms 11-13 have been transformed into product atoms 11-12. We have not made any attempt to specify, for example, that atom 11 in the reactant reaction site corresponds to atom 11 in the product reaction site: such mappings may only be made if assumptions are made as to the mechanism of the reaction. As advocated by Hendrickson(35) we are only concerned with the overall structural changes that have taken place.

If we consider the methyl groups present in the reacting molecules, the possible mappings are, in an obvious notation,

$$1, 6 \longleftrightarrow 1, 6$$

and we wish to detect the equivalences

$$1 \longleftrightarrow 1 \quad \text{and} \quad 6 \longleftrightarrow 6.$$

Equivalent atoms within a single molecule may be detected by application of the Morgan algorithm(141). This partitions the atoms present by considering the number of their attachments, the first order connectivity; as connectivity values rarely exceed four or five, furthur refinement is obtained by consideration of higher order connectivities. The nth order connectivity of an atom is calculated by summing the (n-1)th connectivity values of all the adjacent atoms; thus the two reactant methyl groups of Fig. III.1 may be differentiated by their third order connectivities since their sets of adjacent atoms have different surrounding bond patterns. The discriminatory power of the procedure may be furthur increased by the use of additional properties, such as atom type and the surrounding bond orders, in conjunction with the connectivity (143); at the same time, the nth order property (connectivity) value of an atom a_i, $V^n_{a_i}$, more accurately represents a circular substructure of radius (n-1) bonds centred upon a_i.

We may consider the number $V^n_{a_i}$ to be a hash of its parent circular substructure which may be obtained without a detailed atom by atom investigation of the feature that it describes. Hash coding, or content addressing, is a filesize-independent method of table search which has been widely used for dictionary lookup using alphanumeric character strings(144,145) but it has also been used for chemical structure handling. Early versions of Feldman's substructure search system(146,147) used a hash of the molecular formula and Feldman also mentions that hashing is used extensively in the name file of the CAS Registry System; more recently, entire connection tables have been used as the hashing algorithm's source string(148) and similar work has been reported by Wipke et al(149), Evans et al. used a topological index which could be considered as a hash of a connection table(150) and an analogous approach has been described by O'Korn(151) and Freeland et al.(152). All of these workers were, however, interested in obtaining search codes for registration rather than for substructral representations; the closest approach to the present work, and that described in the next chapter, would appear to be that of Dubois (153).

The matching procedure is based on two principles. Firstly, a modification of the Morgan algorithm is applied simultaneously to both of the reacting molecules so that inter-, rather than intra-, molecular equivalences are detected. Secondly, we assume that the nth order property value $V^n_{a_i}$ is a unique representation of an (n-1) bond radius, circular substructure centred upon atom a_i. Hence if $V^n_{r_i} = V^n_{p_j}$, the reactant and product atoms r_i and p_j may be considered to be at the centre of identical substructures and these areas may be assumed to be the same without a detailed examination of the constituent atoms, i. e. an isomorphism is presumed to exist.

For the initial property value of each atom, $V^1_{a_i}$, we have used an integer derived from the atomic 'dot-plot' symbol which uniquely describes the type and bond pattern of a wide range of atoms(109,54). Higher order property values are obtained from the equation

$$V^n_{a_i} = 2V^n_{a_i} + \sum V^{n-1}_{a_j}$$

where the summation is over all the atoms j that are adjacent to a_i. The function is similar to that used by Shelley and Munk to identify intra-molecular atomic equivalences(155). Higher order property values are calculated for all of the reactant and product atoms until there are no remaing pairs for which $V^n_{r_i} = V^n_{p_j}$. At this stage, the pair(s) of atoms for which $V^{n-1}_{r_i} = V^{n-1}_{p_j}$ are noted and all the atoms within (n-2) bonds of these atoms, the 'match radius' r, deleted from the reactant and product connection tables. The partitioning of the sets of reactant and product atoms is similar to the use of atomic properties of various kinds in the set reduction techniques first proposed by Sussenguth and Unger (116,115) for the detection of (sub)graph isomorphisms. These procedures involve the generation of pairs of corresponding subsets of the nodes present in the trial and query structures; these subsets are subsequently partitioned by the application of a range of characteristics to determine correspondences between the individual nodes in the two structures. The properties used for the partitioning include atom type and degree of connectivity though both authors point out that these may not be sufficient in some cases. More recently, Figueras(124) has considered higher degrees of connectivity and a similar approach has been described by Schmidt and Druffel(156). The partitioning procedure means that in many cases an isomorphism, or the lack thereof, may be detected without the need for any iterative atom by atom searching(157,158). Our application goes one step furthur insofar as substructures,

rather than single atoms, are matched without such a search.

Applying the procedure to the reaction of Fig. III.1, possible matches, $r_i \longleftrightarrow p_j$, are obtained until the sets of $V^6_{a_i}$ values have been calculated at which point no mappings remain for which the property values are the same. We hence obtain the equivalence

$$1 \xleftrightarrow{r = 4} 1$$

which, after the elimination of all the atoms contained within the match radius, results in the reaction diagram shown in Fig. III.2: the atoms that are shaded in the Fig. are those which have been deleted, i. e. noted as not being involved in either of the reaction sites. Consideration of the remaining atoms yields the equivalence

$$10 \xleftrightarrow{r = 3} 10.$$

No furthur mappings can be found, so after updating the reactant and product adjacency matrices, the procedure terminates to yield the reaction scheme shown in the lower half of Fig. III.3.

The actual implementation has three additional features which should be mentioned. Firstly, we must allow for multiple equivalences as illustrated by the reaction shown in Fig. III.4 for which we obtain the mapping

$$14,\ 15,\ 16 \xleftrightarrow{r = 7} 14,\ 15,\ 16.$$

The atoms, and the substructures centred upon them, may be deleted only if all three reactant atoms have the same set of three possible matches, i. e. if the members of the reactant set are equivalent one to another as well as to the product set. If this is found to be so, an arbitrary assignment is made for each member of the reactant set and the deletion process then takes place as normal. Secondly, we have defined a minimal match radius of two bonds: early work showed that if a match radius of one is permitted, corresponding to the matching of two augmented atoms(108,159), there is a slight,

but noticeable, increase in the number of mappings which, although representing isomorphic substructures, do not correspond in chemical terms. Finally, for a match radius r, only the atoms within (r-1) bonds are deleted. This step is taken to guard against cases such as the reaction of Fig. III.5 where the bonds attached to the outermost atoms, r_4 and p_4, are differently oriented in the two structures. These latter two restrictions tend to reduce slightly the number of atoms eliminated; thus the reaction of Fig. III.1 is now analysed as shown in Fig. III.6 with the adjacent carbon being included in the reaction site. It should be noted that, in some cases, these limitations may extend the derived reaction sites quite considerably: thus for the reaction shown in Fig. III.7, the methyl groups attached to the tetravalent carbon atoms are all noted as being in the reaction site.

A further example of the method of analysis is shown in Figs. III.8 and III.9. The maximal mapping obtained for the reaction shown in the upper portion of Fig. III.8 is

$$7 \xleftarrow{r = 5} 7$$

and deletion of the appropriate substructures gives the reaction diagram shown in the lower half of the Fig.. The next mapping arises from the two remaining phenyl groups and is

$$14,\ 23 \xleftarrow{r = 4} 14,\ 23$$

which produces the furthur deletions shown in the upper half of Fig. III.9. The final mapping is

$$1 \xleftarrow{r = 3} 1$$

which gives rise to the final analysis shown in the lower half of the Fig..

Having given two examples of the procedure, we close this section with a brief description of the basic algorithm: a detailed implementation is given in Appendix III.

(i) read reactant and product connection tables.

(ii) m := 1; assign all $V^1_{r_i}$ and $V^1_{p_j}$ values using the units values.

(iii) generate higher order property values until there are no atom sets for which $V^m_{r_i} = V^m_{p_j}$ (m = n).

(iv) determine the most similar atom pairs(r_i, p_j), i. e. those sets of atoms which obey the relationship $V^m_{r_i} = V^m_{p_j}$ $(1 \leq m \leq n-1)$.

(v) n := n - 1.

(vi) if there are no unique mappings go to (vii) else delete all atoms within an (n-1) bond radius of the atoms r_i and p_j for all the pairs of atoms (r_i, p_j).

(vii) if there are any remaining multiple mappings, assign equivalent reactant and product atoms and then go to (vi).

(viii) determine the most similar remaining atom pairs, set n accordingly and then if $n > 2$ go to (v).

(ix) output connection tables with the atoms in the reaction site suitably tagged.

III.2 Results of the procedure

The algorithm was tested using a sample file of 534 reactions taken from the WLN data base described above. The WLNs of the reacting molecules were converted to Crossbow connection tables (16,109) using software kindly provided by ICI Ltd.(Pharmaceuticals Division) and these tables were then used to produce redundant adjacency matrices using a program written by the author(a listing of which is included in Appendix III). An advantage of the Crossbow representation is that the "units" section identifies both the atom type and the bond pattern around each atom in a molecule within a single character, the dot-plot symbol(154). The binary representations of these symbols were used as the first order property values for the structure matching procedure which was implemented in an ALGOL 68R program(160) which was run on the University ICL 1906S computer. The program contained *circa* 200 lines of code and occupied 35K words of core storage; it required 63 cpu seconds, exclusive of transput, to process the file of 534 reactions, that is between eight and nine reactions per second. The input to the program consisted of the redundant adjacency matrices together with a vector giving the units values of the atoms; the output was identical except that the vector now contained a note as to whether a given atom was to be considered as being part of the reaction site. The results of this computer run are shown in Fig. III.10. Inspection of the output showed that successful analyses were obtained for 491 of the reactions(92%) where a satisfactory analysis is judged to be one that adequately represents the change though, as noted above, additional atoms may be included in the reaction site. With this proviso, the analyses exhibit a quite striking degree of reaction site localisation; examples to justify this statement are shown in Figs. III.11 to III.16

which contain the reactions, the mappings and the derived reaction sites. For large match radii, the property values may become very large and the reactions described under "Overflow" in Fig. III.10 correspond to cases where the values became too large for the computer word reserved for them: in both cases, the match radius , if calculated, would have been over 20.

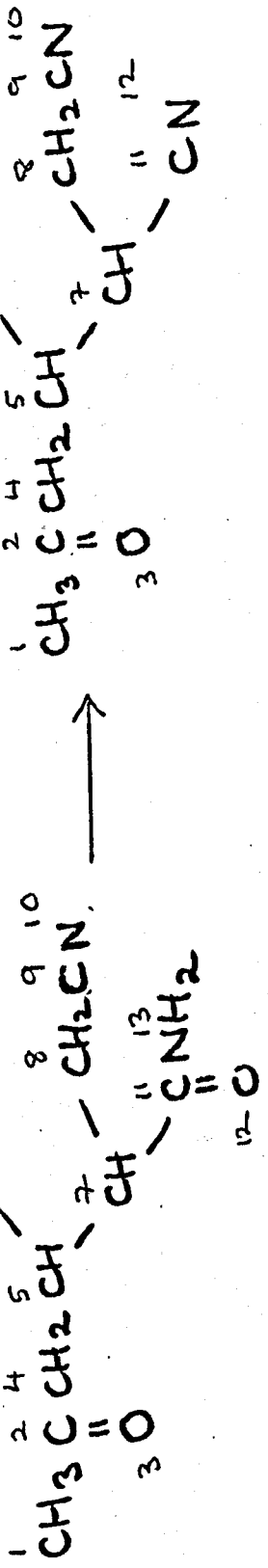
Reactions where no atoms were deleted are shown in Fig. III.17 and III.18. The two reactions of Fig. III.17 are not processed since an ambiguous mapping is obtained, i. e. more than one reactant atom maps onto the same product atom or vice versa. For symmetric molecules an arbitrary assignment procedure could be invoked to overcome this problem; Ming and Tauber(123) give a detailed description of assignment procedures for structure matching using a backtrack procedure(161) but these are not generally applicable here since incorrect assignments could not be detected subsequently. We will also not obtain any mappings if the reaction has occurred in a fairly small molecule, e.g. the reactions of Fig. III.18 where no pairs of atoms have a match radius greater than 1.

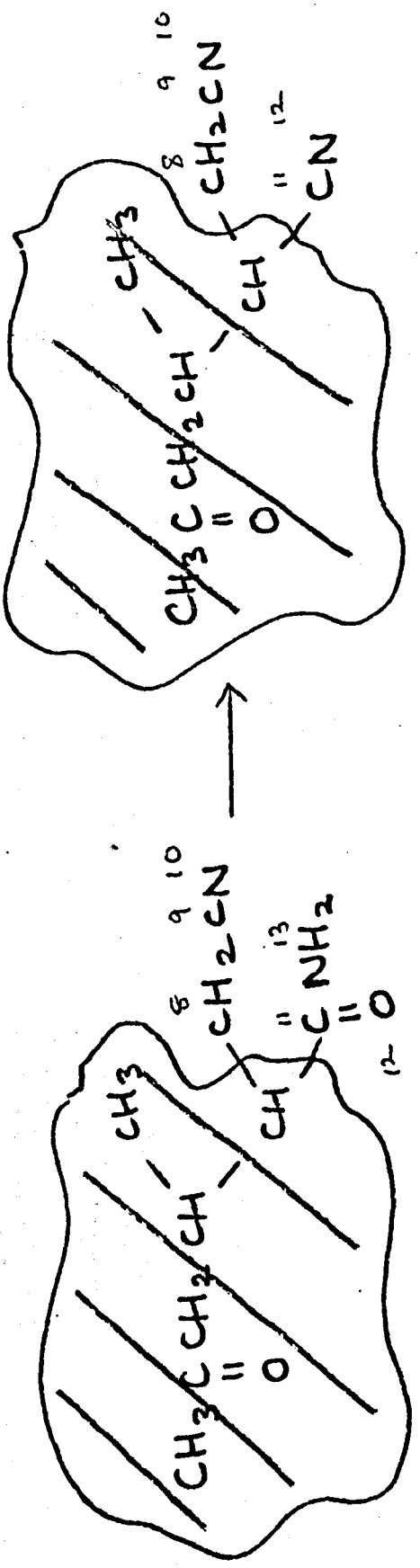
Some of the failures are shown in Figs. III.19 to III.23 and it will be seen that they arise for a variety of reasons. Problems arise in the first example due to the shift of the allyl group. The second reaction, Fig. III.20, is one of the few undetected cases which contradict the assumption that equal sets of property values correspond to identical substructures: the reaction involves a functional group shift not detected by the matching algorithm. Incorrect mappings will be obtained if an atom involved in the change is matched with a non-reacting atom; thus in Fig. III.21, the invalid equivalence

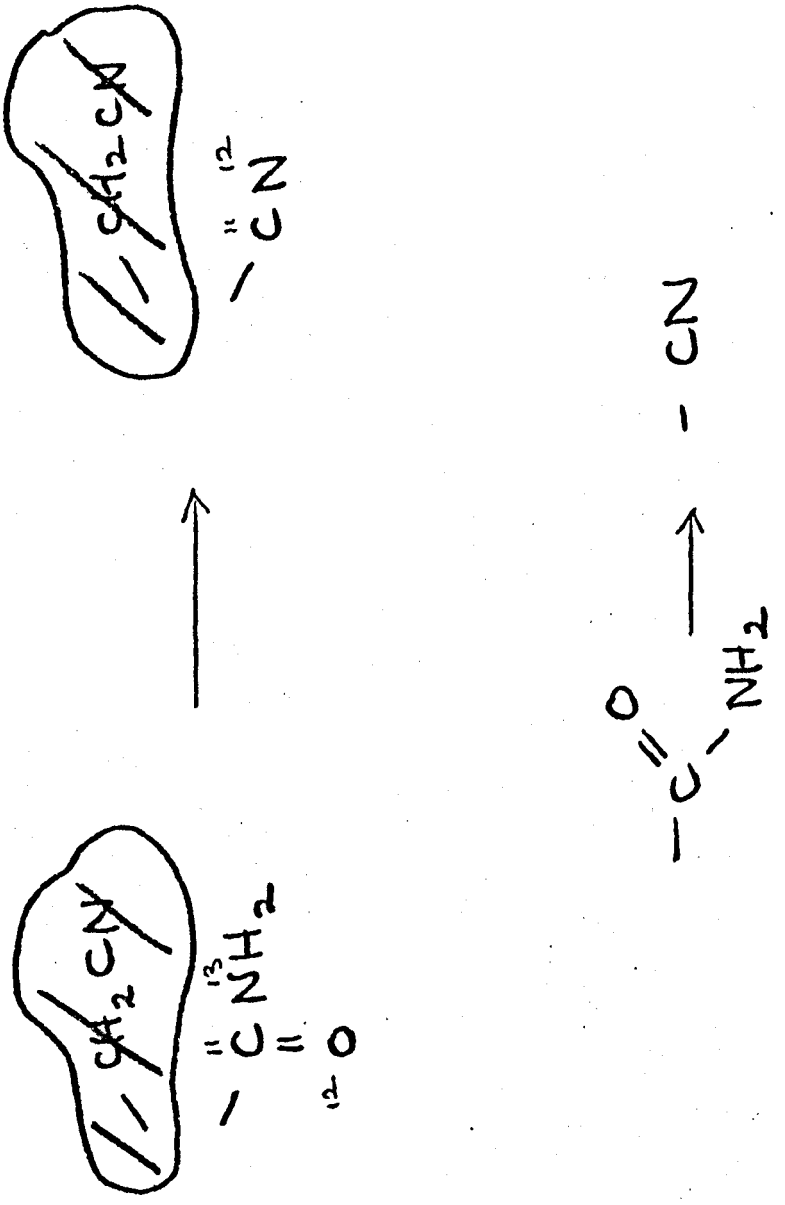
$$16 \overset{r = 5}{\longleftrightarrow} 25$$

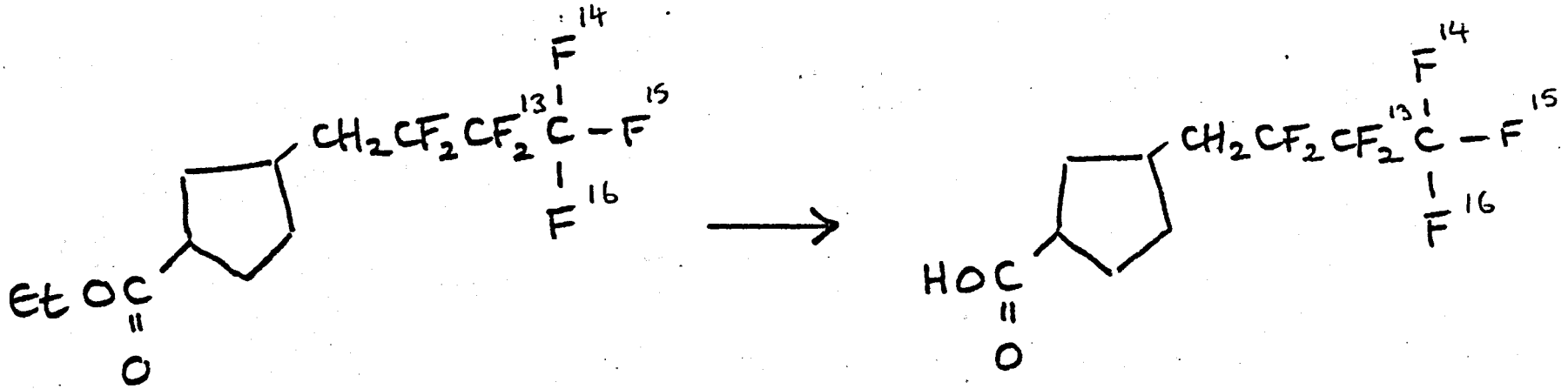
is obtained. Other failures are shown in Figs. III.22 to III.23.

It will be seen from Fig. III.10 that the number of undetected failures is very small indeed. A graph isomorphism routine could be incorporated after a mapping has been detected to check that the circular substructures that have been mapped are in fact isomorphic and this would probably be quite fast in operation since it is generally a simpler matter to prove that an isomorphism does not exist between two graphs than to prove the contrary(162). This, however, would only be useful for the one or two reactions in which a non-isomorphic mapping is obtained: most of the failures, on the other hand, arise from correctly identified isomorphisms which, however, do not correspond with what has taken place in chemical terms. The assumption that equal sets of property values correspond to identical substructures would seem to be valid for the overwhelming number of reactions considered. Note, however, that such an assumption probably would be much less applicable to a general graph matching algorithm in a substructure serach system where a wide range of structures are to be matched against the query: in the present application, it may be taken a priori that a large degree of similarity exists between the two (sets of) molecules being considered and hence if a mapping is found, it is almost certainly a valid one.

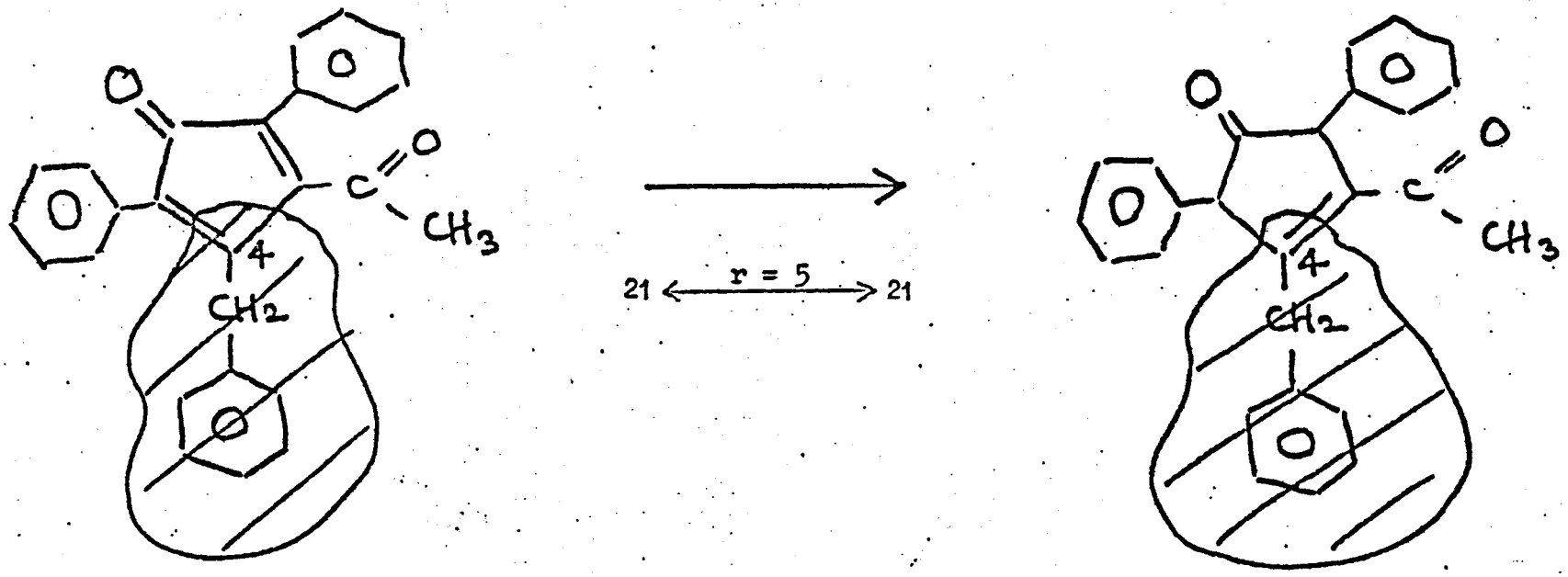



CH3
CH3
CH
CH2
C
=O
CH
CH2CN
=C
NH2
=O
CH3
CH3
CH
CH2
C
=O
CH
CH2CN
=C
N

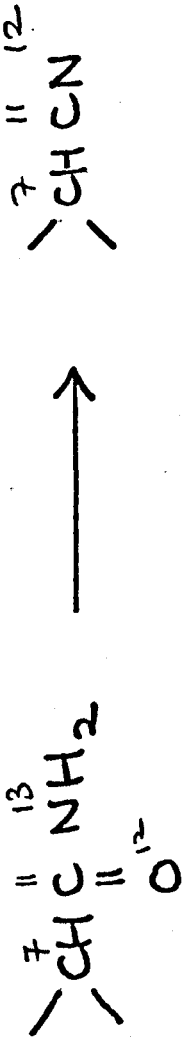

CH2CN
CN
CH2CN
C
NH2
O
CN
C
NH2
O

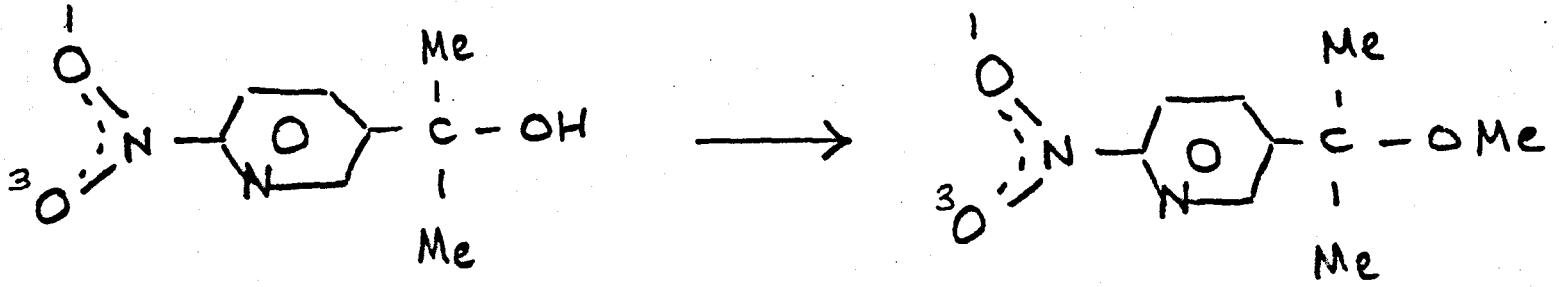


Example of a reaction containing multiple equivalences. Note that, from hereon, only certain of the atoms will be numbered to avoid cluttering the diagrams.



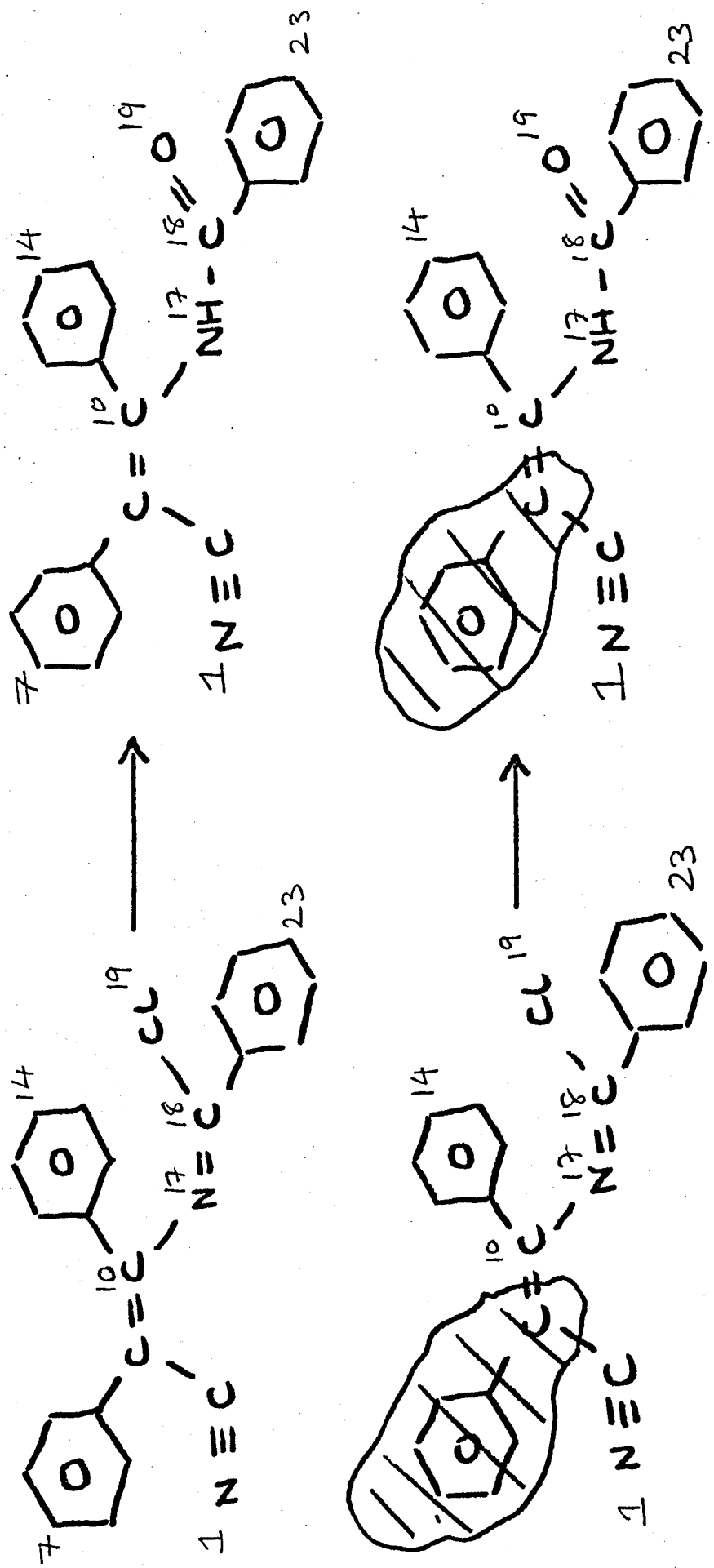
An incorrect mapping is obtained due to the differing orientations of the bonds around r_4 and p_4.

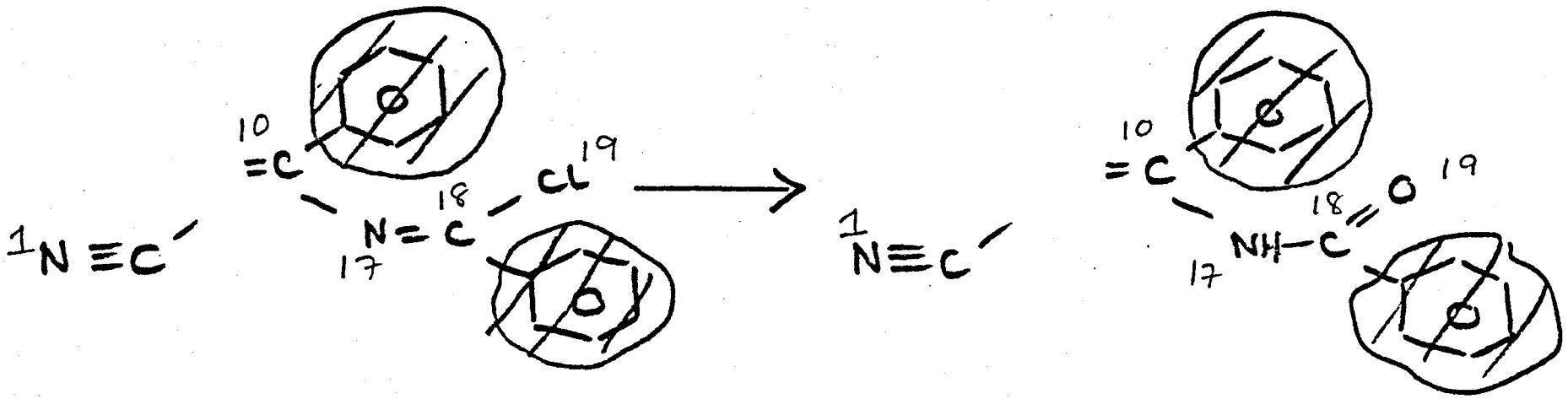


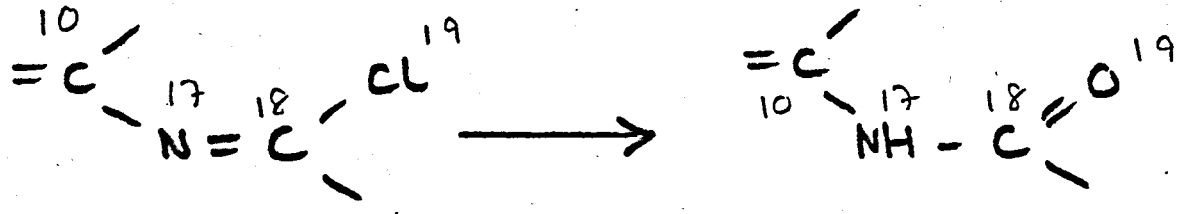


1, 3 $\xleftarrow{r = 6}$ 1, 3

$-CMe_2OH \longrightarrow -CMe_2OMe$

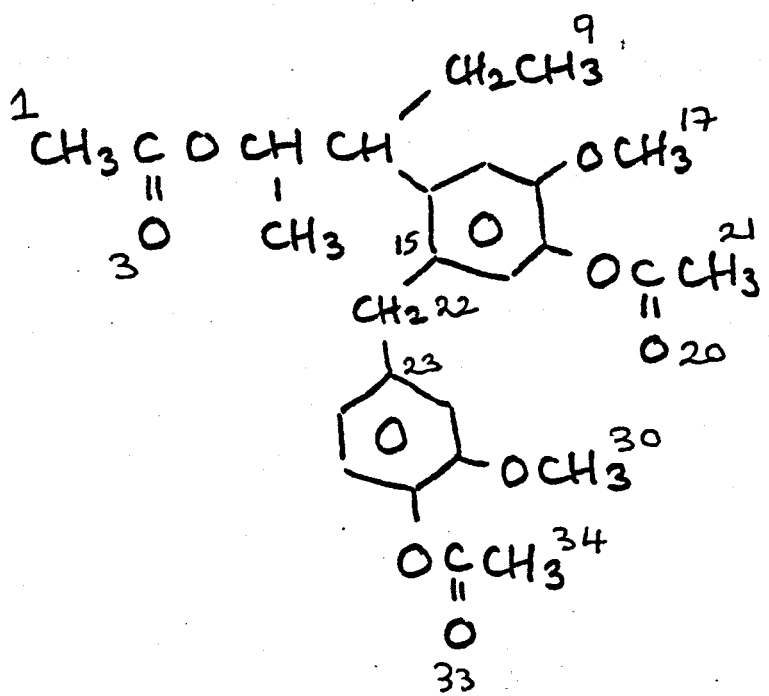


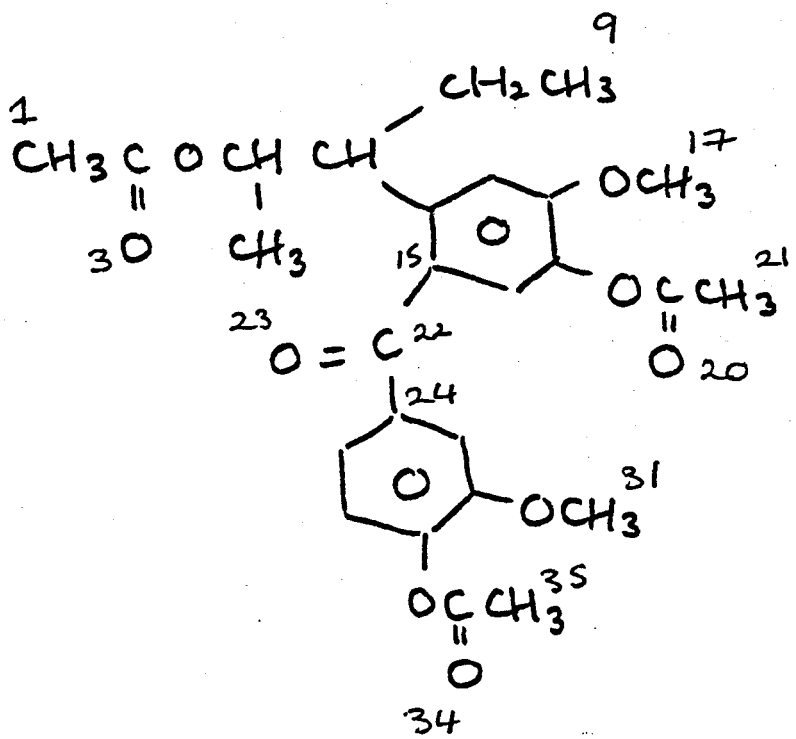




Successful analyses	491
Reactions for which overflow took place	2
No matches obtained	30
Incorrect mappings made	11

Results for analysis of sample file of 534 reactions using the graph matching algorithm of Chapter III.





1 ⟷ r = 6 ⟷ 1

3 ⟷ r = 6 ⟷ 3

33 ⟷ r = 6 ⟷ 34

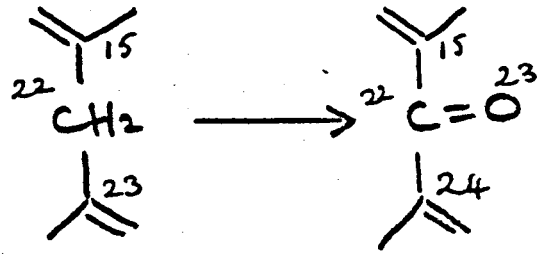
34 ⟷ r = 6 ⟷ 35

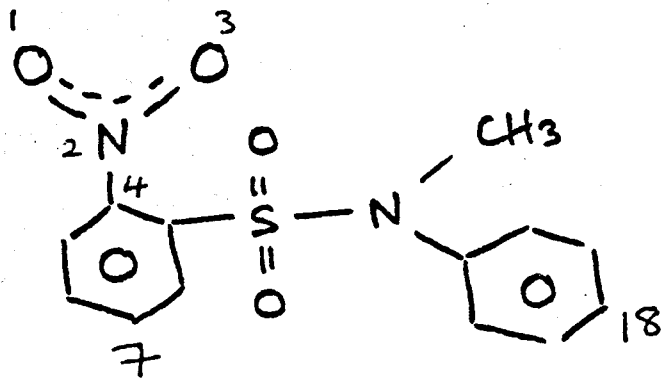
17 ⟷ r = 5 ⟷ 17

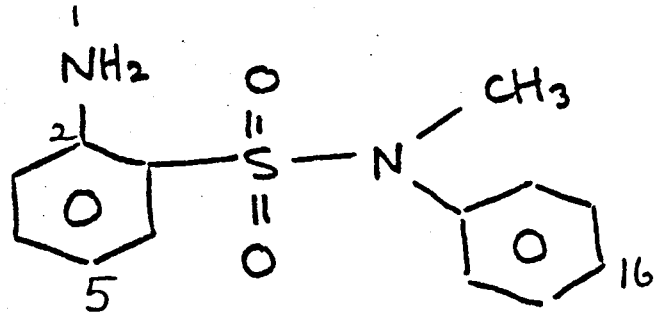
21 ⟷ r = 5 ⟷ 21

9 ⟷ r = 4 ⟷ 9

30 ⟷ r = 4 ⟷ 31

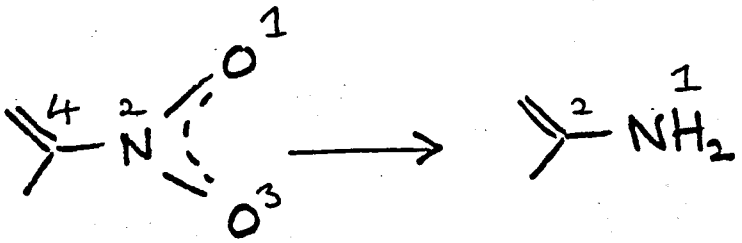


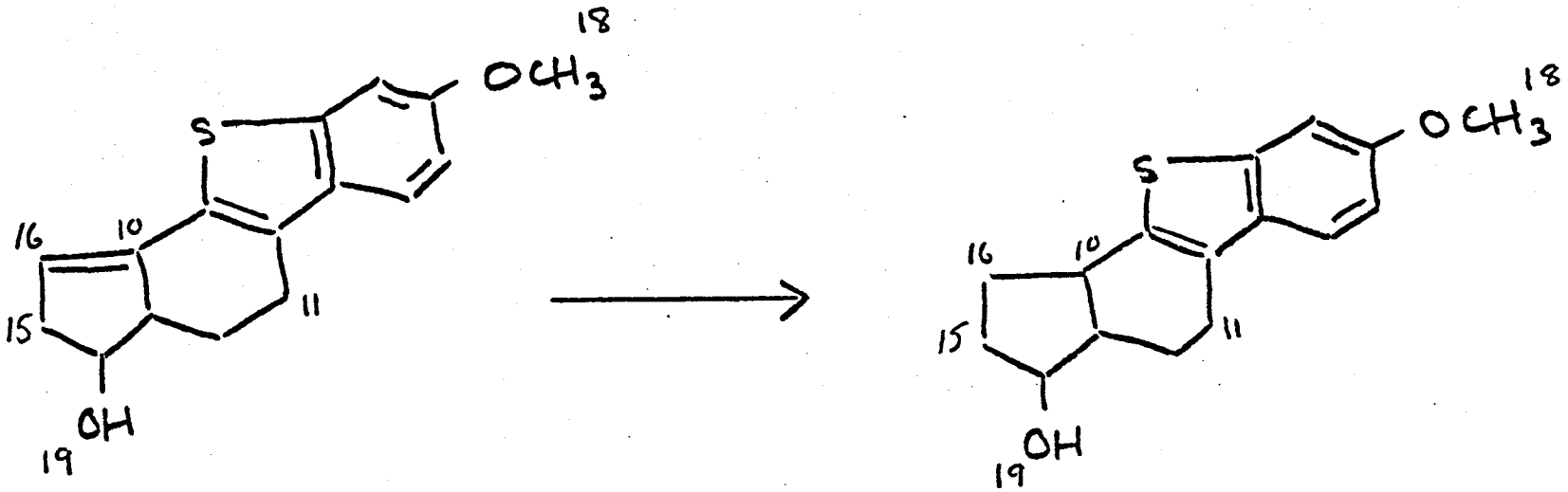




18 ⟷ r = 7 ⟷ 16

7 ⟷ r = 3 ⟷ 5

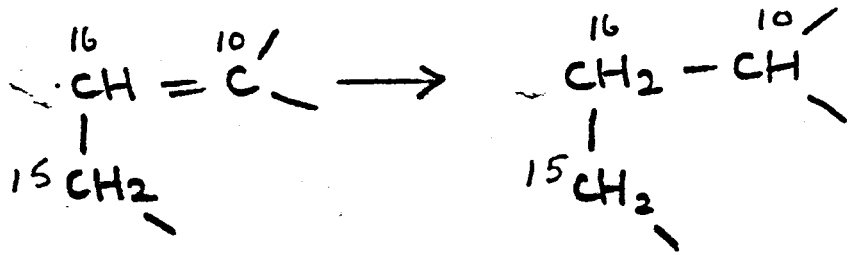


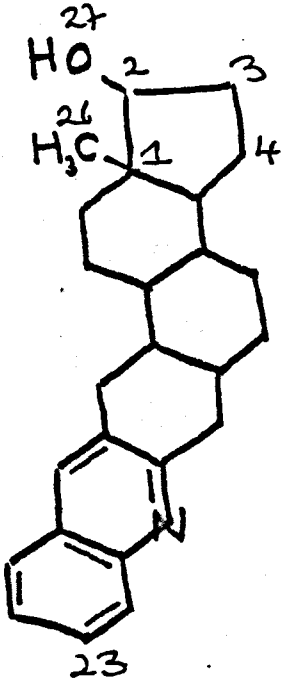


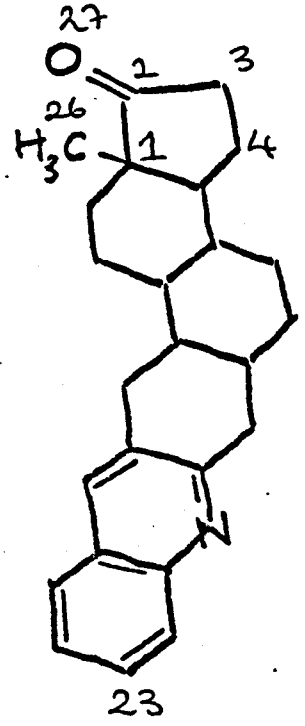
18 $\xleftarrow{r = 6}$ 18

11 $\xleftarrow{r = 2}$ 11

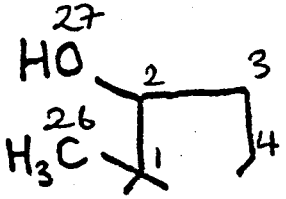
19 $\xleftarrow{r = 2}$ 19

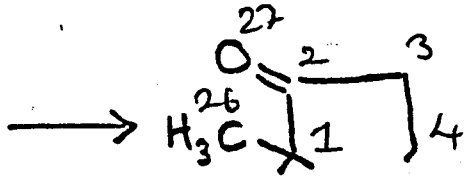


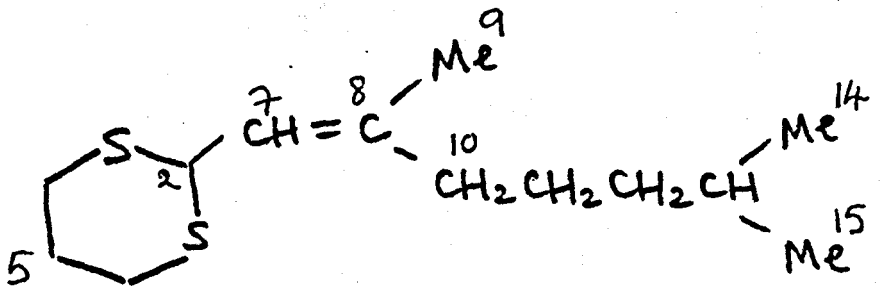

27
HO
2
3
26
H3C
1
4
23

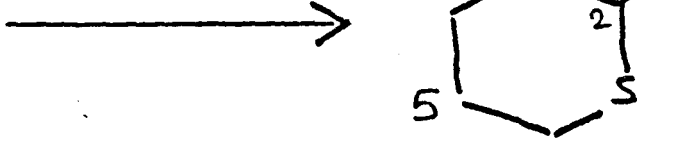

27
O
2
3
26
H3C
1
4
23

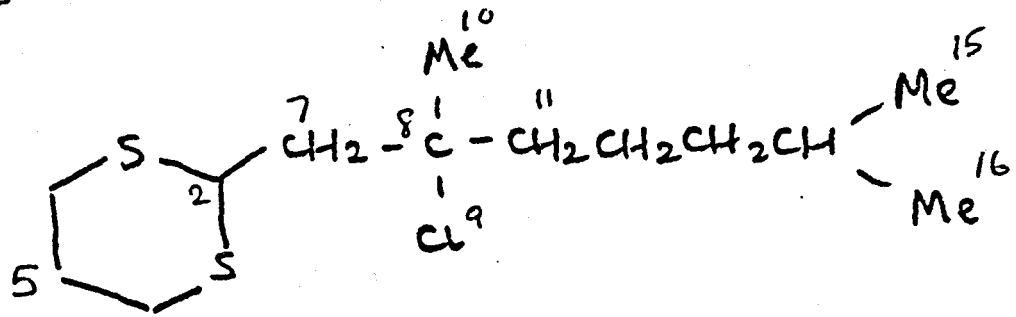
23 ⟷ r = 11 ⟷ 23


27
HO
2
3
26
H3C
1
4


27
O
2
3
26
H3C
1
4

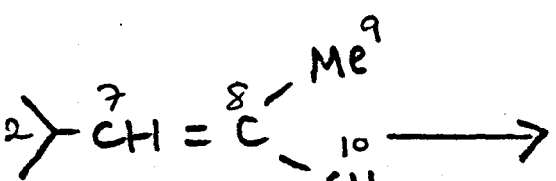


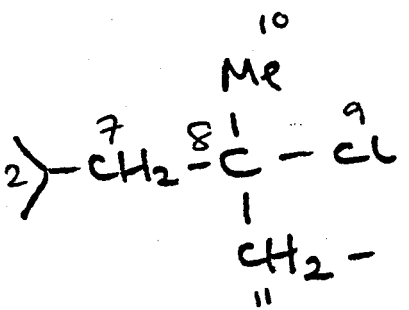


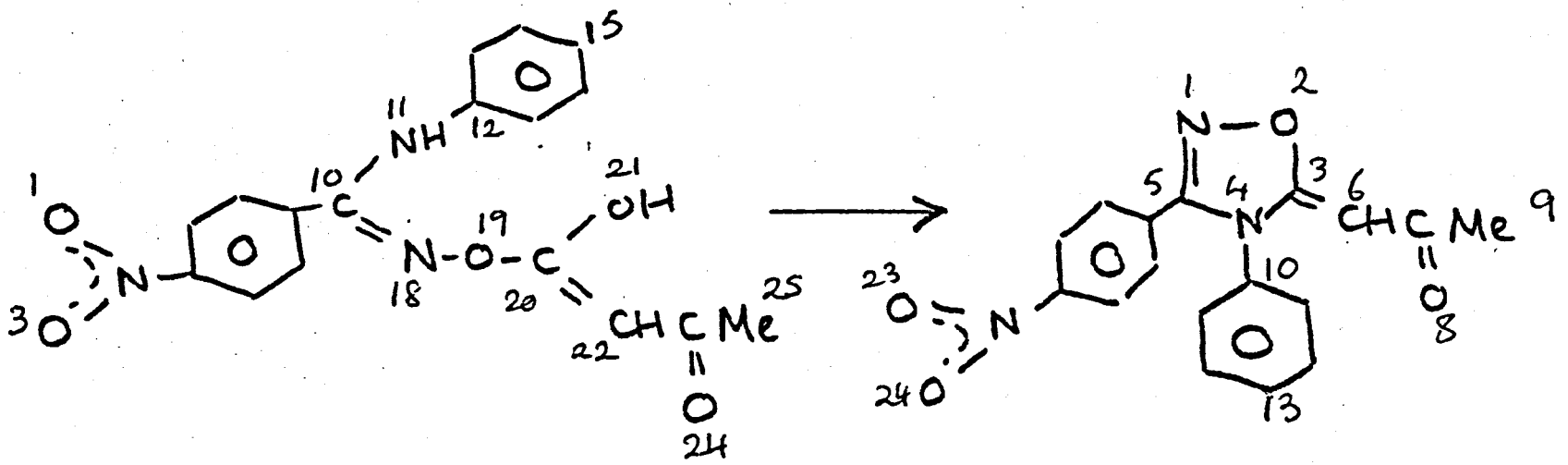


14, 15 $\xleftarrow{r = 4}$ 15, 16

5 $\xleftarrow{r = 3}$ 5





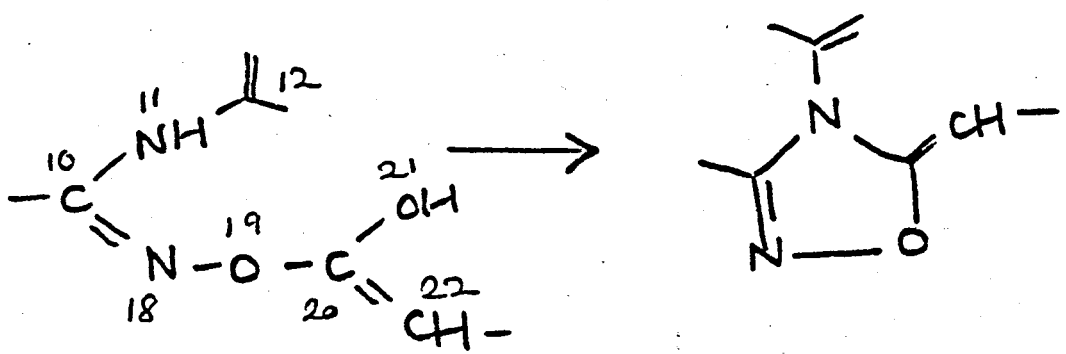


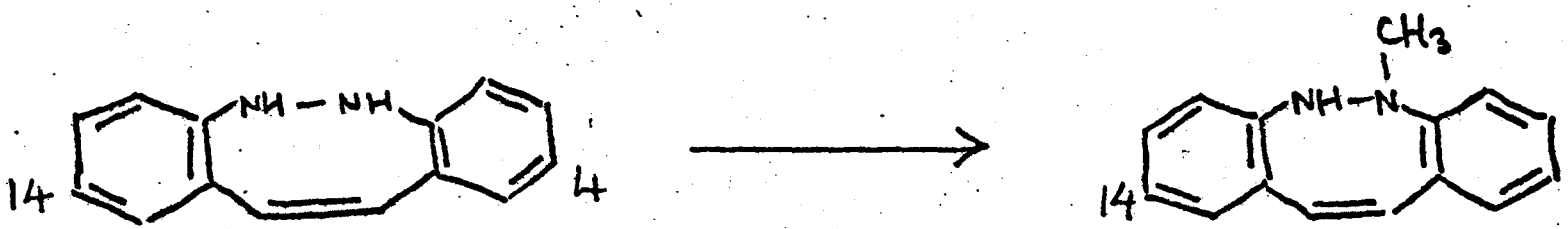
1, 3 ⟷ r = 5 ⟷ 23, 24

15 ⟷ r = 3 ⟷ 13

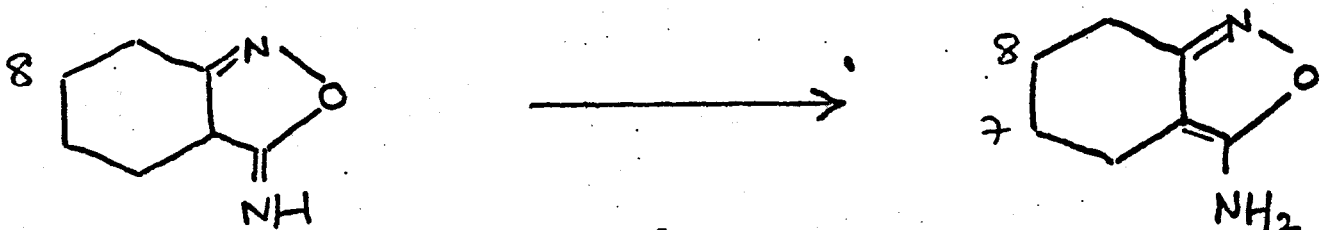
25 ⟷ r = 2 ⟷ 9

24 ⟷ r = 2 ⟷ 8



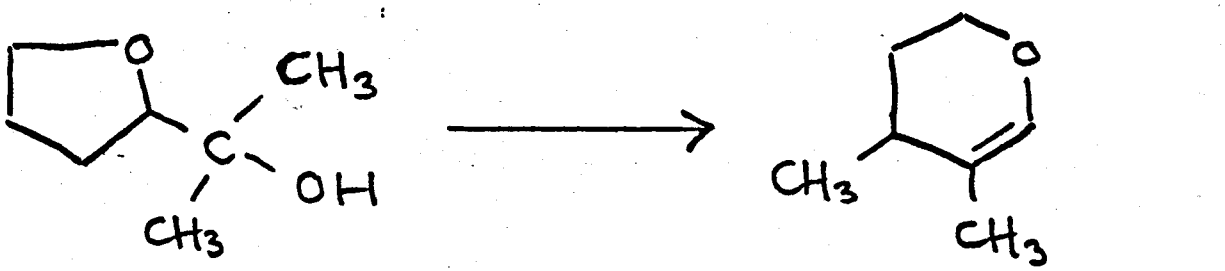


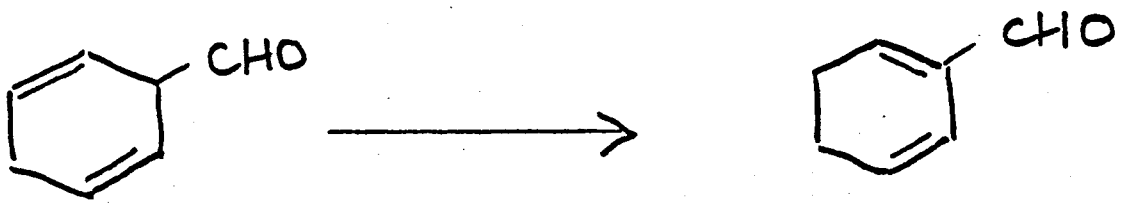
4, 14 $\xleftarrow{r = 4}$ 14



8 $\xleftarrow{r = 2}$ 7, 8

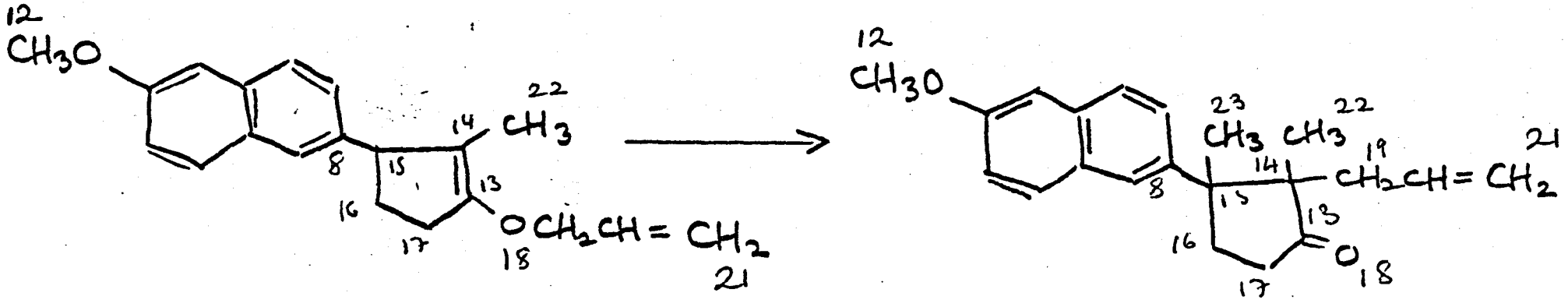
Examples of duplicate mappings.



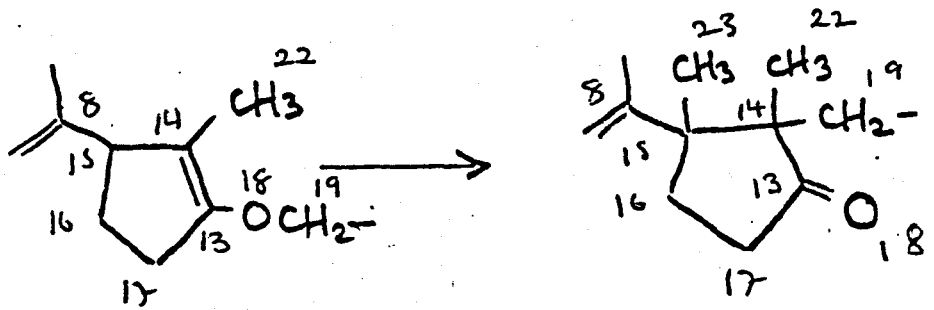


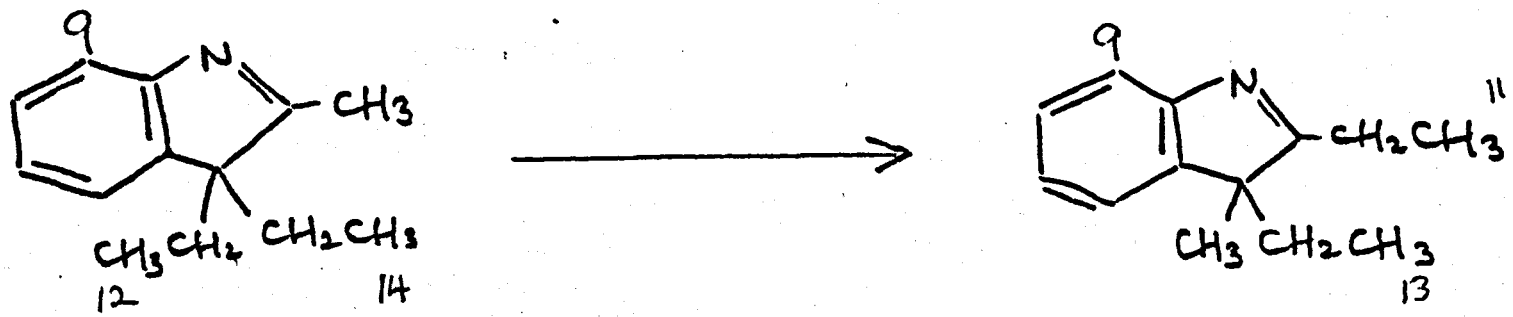
Reactions in which no mappings are identified

due to the small size of the reacting molecules.



12	r = 7	12
21	r = 2	21

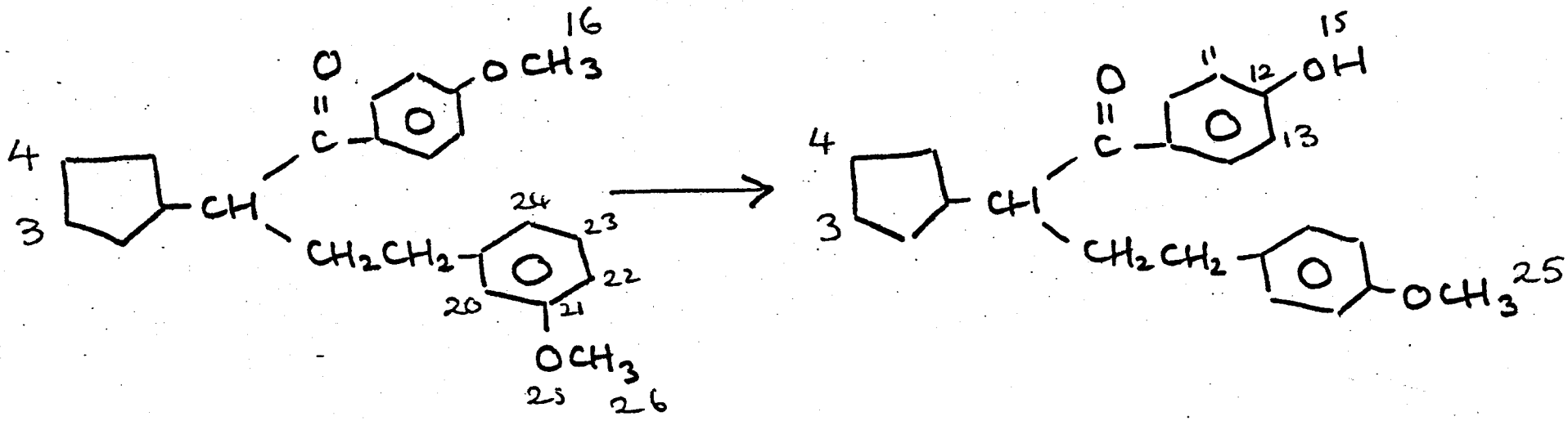




9 r = 5 9

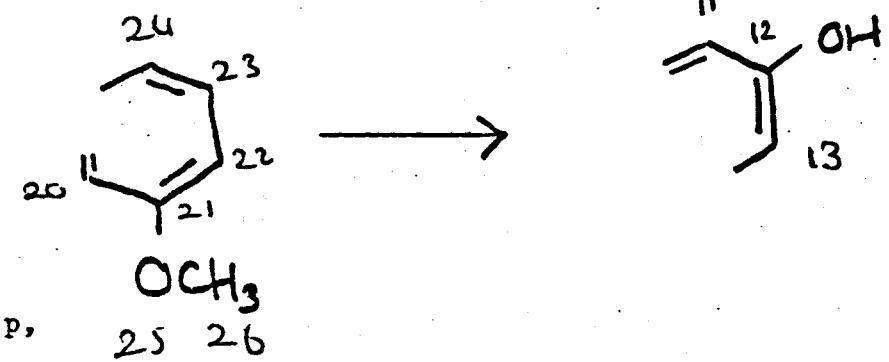
$$CH_3^{-(12)}, CH_3^{-(14)} \longrightarrow CH_3^{-(11)}, CH_3^{-(13)}$$

Note added in proof: a small routine has been added to the structure matching program to match any unconnected atoms remaining at the end of the analysis which results, in the present case, in the elimination of all of the reactant and product atoms and the analysis is hence rejected as being invalid.

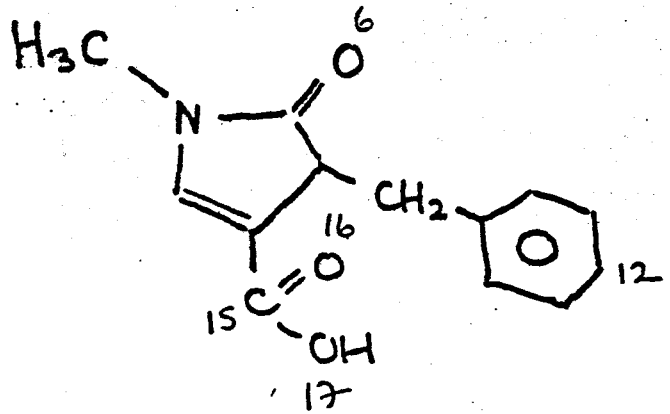


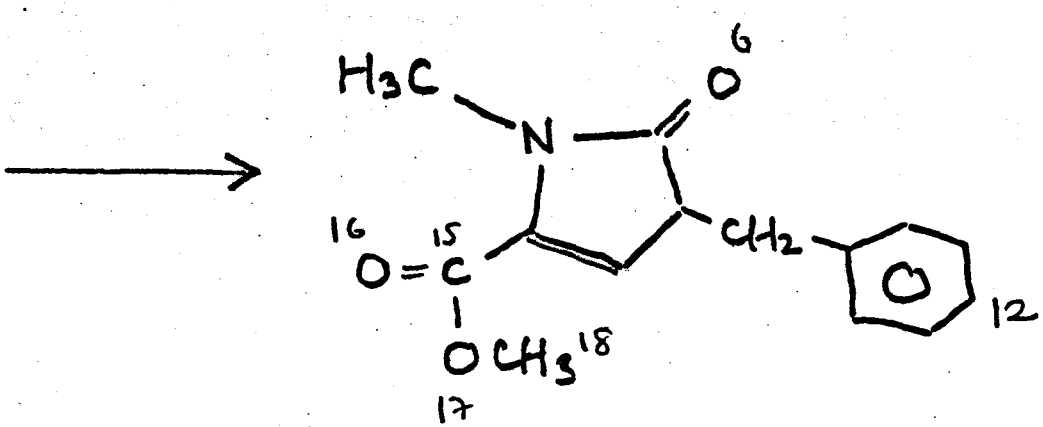
3, 4 $\xleftarrow{r = 7}$ 3, 4

16 $\xleftarrow{r = 5}$ 25



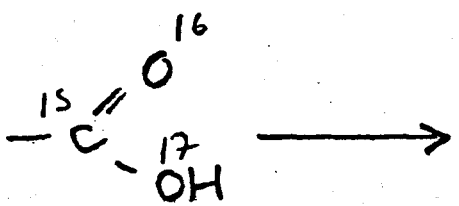
A possible check for failures of this sort could be easily implemented: knowing the total numbers of atoms, r and p, in the reactant and product molecules, the difference in size between the two reaction sites should be r-p atoms if a consistent set of mappings have been obtained.

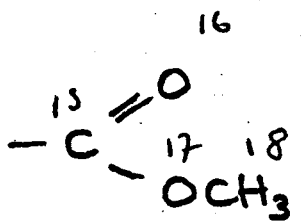


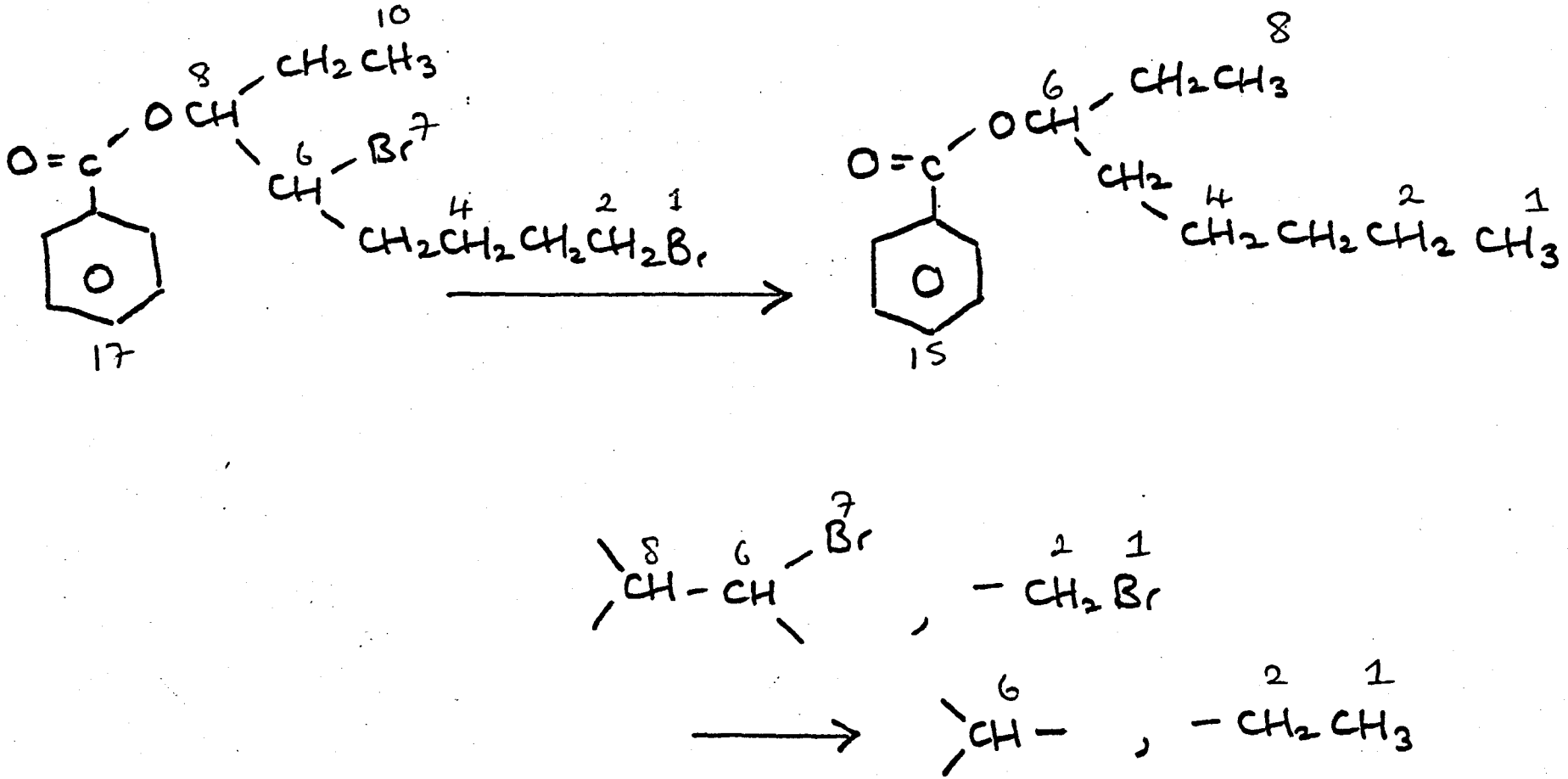


12 ⟷ 12 (r = 5)

6 ⟷ 6 (r = 4)







CHAPTER IV

The automatic generation of screen sets for chemical substructure search systems.

IV.1 Introduction

The rapid build-up of large machine-readable files of chemical compounds has led to a need for sophisticated search procedures to meet the needs of users for improved access to structural and related property data. Searches may be performed both for individual compounds and for groups of molecules possessing certain substructural features in common. The first of these tasks, registration, is carried out to ensure that there is no scattering of the information pertaining to a single compound and that information concerning different compounds does not become confused; general descriptions of the problems involved are given by Ash(16) and Evans(163) and a specific implementation by Evans et al.(150).

Substructure search is normally carried out using a multi-level approach in which increasingly specific search techniques are applied to rapidly diminishing sections of the structure file. The search strategies adopted will be in large part dependent upon the structural representation used in the file and questions may therefore be answered by nomenclature-based or structural fragment codes(59,164), by string searching linear notations(103,165) or by searching of bit screens generated from notations(166), nomenclature(58) or connection tables(125). There will be some questions, however, which can only be adequately answered by a comprehensive atom-by-atom search for which a connection table record, of some sort, is essential. Such searches are equivalent to detecting the isomorphism of a subgraph, the query structure, with a complete graph, a compound from the structure file; subgraph isomorphism detection is known to belong to the class of problems known as NP-complete(121,167) for which no efficient algorithms are known and thus may require uneconomically large amounts of computer time if many trial structures need to be matched against the query. Atom-by-atom searching of large compound

files is accordingly feasible only if the number of searches can be kept to a minimum by the rapid and inexpensive elimination of that large portion of the file which cannot possibly meet certain minimal requirements in the query formulation. The structural characteristics which are used to carry out this file partitioning are called screens; we use the term 'screen set' to describe the group of characteristics which are chosen for this purpose.

What criteria should be used in generating a screen set from the astronomical number of possible fragments that could be employed?(168,169). Ray and Kirsch(158) pointed out that the members of a screen set should be independent of one another and be applicable to the whole range of questions that might be expected. In a non-chemical context, Mooers(170) noted that in an ideal situation a set of twenty characteristics, each of which was independent of all of the others and occurred in one half of a file of one million items should be capable of uniquely identifying a single record. Although such requirements are not obtainable in practice, these broad guidelines have played a large part in the development of information systems in general and of chemical structure systems in particular. It is found that there are two factors which prevent screen sets from achieving a near ideal performance; the first is the extremely disparate occurrence of structural features and the second the presence of strong inter-fragment dependencies. In the next section, we discuss previous work dealing with these two factors and follow this by the presentation of an algorithmic method for screen set generation which takes the factors into account.

IV.2 Theoretical considerations in the design of screen sets

It is found empirically that a variety of hyperbolic, and other long-tailed, distributions characterise the behaviour of many parts of library and information systems(171,172,173,174) and in a series of papers(92,108,130,175), Lynch and his co-workers demonstrated that there is an inverse relationship between the frequency of occurrence of a substructural feature and its rank when the features are ranked in decreasing frequency order. An extreme example is provided by the distribution of element types in a structure file: analysis of almost 30000 compounds drawn at random from the Chemical Abstracts Registry System showed that the occurrence of the most frequent atom, carbon, was almost 1000 times that of the tenth-ranked atom, iodine. Moreover carbon, oxygen and nitrogen together accounted for almost 95% of the non-hydrogen atoms in the sample file. The obvious implication is that the value of the element type alone as a screen is highly variable, since searches involving iodine will be highly selective whereas queries involving carbon or nitrogen will require many other characteristics to be specified if an enormous, low precision output is to be avoided. Analysis of larger, more detailed fragment types yields the same broad conclusions but with two qualifiers: firstly, the frequency difference between differently ranked fragments is much reduced but, secondly, the totality of fragment types, the variety, is much increased with the majority of types occurring very few times indeed. An additional problem is that the queries addressed to a structure file are a fair representative of the file's contents(164,176) and it will hence be the very frequent features which are most often specified at search time.

The disparate frequencies of the fragment types may be compensated for in the design of a screen set by employing

varying levels of description, the frequently occurring characteristics being delineated in some detail whilst the less common features are described in more general terms. In this way, we may achieve a balance between the proliferation of low incidence fragments of superfluous specificity and the small number of high incidence, low precision fragments. At the same time, the occurrences of the resultant screen set members will become much less disparate than if a single level of description were to be employed. This move towards screen equifrequency is, however, lessened by the necessity of describing frequent characteristics at the more general levels, as well as in detail, to allow of easy query encoding since, otherwise, the union of many highly specific features may be required in order to describe a more general feature common to all of these.

A theoretical justification for such an approach is obtained from simple considerations of information theory. Shannon's mathematical theory of communication(177) considers the statistics of symbol occurrences in messages and gives a quantitative measure of the maximum amount of information that may be gained from a message encoded using a given collection of symbols. The actual nature of the symbols is immaterial: thus they may be the letters of an alphabet, the fragments in a screen set or the indexterms in a controlled vocabulary indexing language. The theory shows that the average amount of information conveyed by each symbol is given by

$$H = -\sum_{F=1}^{N} Q(F)\log Q(F) \qquad (1)$$

where H is called the entropy of the symbol set and the values $Q(F)$ are the probabilities of occurrence of each symbol F in the set of N symbols. The probabilities may be approximated by the relative occurrences, $N(F)$, of the symbols in the set, i. e.

$$Q(F) = N(F)/N \qquad 0 \leq Q(F) \leq 1 \qquad (2)$$

Irrespective of the nature of the symbols in the set, the entropy will be at a maximum when the symbols are equiprobable, i. e. they occur with equal frequencies, and independent of one another; in such a case, the entropy, H(max), is simply given by the logarithm of the symbol set size. A rapid measure of the coding efficiency of the set is given by the relative entropy H(R)

$$H(R) = H/H(max)$$

where H is the actual measured entropy. It should be noted that the relative entropy measure becomes somewhat insensitive as its value tends to unity.

An additional problem is that the screenout performance of a screen set cannot be directly calculated from fragment incidence data since it is found that the incidences of the screen set members are not indepedent of each other(178). An analysis of the co-assignment frequencies of pairs of screens showed that the association between fragment types of a given size increased directly as the size and that certain of the screen-pair associations were sufficiently large to have a considerable effect on the performance of a query involving that pair of screens. If a query involves two positively associated screens, the association will reduce the screenout whilst the converse will occur if the screens are negatively associated.

Many of the positive fragment associations may be easily explained in terms of overlap between fragments. The iterative fragment generation method developed at Sheffield(179) considered all possible centres of a given type in a structure, the types being bonds, atoms and rings. Thus, once fragments had reached a certain size, those generated from adjacent centres would start to overlap and the region of overlap would increase with increasing fragment size. Hence if a substructure occurred fairly frequently, fragments derived from it would have quite high,

positive associations, e.g. the carboxylic group yields the simple pairs C-O, C-C and C=O all of which share the same carbon atom. Negative fragment associations, which would improve the screenout, are more difficult to explain and this is also true for both positive and negative associations between individual atom types where no overlap is possible. The study concluded that, in practice, no consideration need be given to fragment associations as long as the screen set members were not too large.

The Sheffield group considered only associations between fragments of the same size, but iterative fragmentation algorithms introduce very strong associations between a fragment and its immediate parent, i. e. the fragment from the previous iteration of the fragmentation algorithm from which it has been derived. It is clear that if the incidence of a fragment is not dissimilar from that of its parent, one of the fragments is redundant and should not be included in the screen set; the filial fragment will have the lower frequency and thus should be deleted to permit easier query encoding at search time.

A model for such associations has been developed by Hodes and been applied to the formation of a screen set in use at the Walter Reed Army Institute of Research(168,180,181). The discrimination of a fragment, D(F), is defined as being the reduction in uncertainty when the fragment is used to partition the file, i. e. is assigned to the appropriate molecules within it. Using the notation given above, application of the fragment will cause the file to be reduced in size from N to N(F) so that the change in uncertainty will be

$$D(F) = \log N - \log N(F) \qquad (3)$$

which, substituting from equation (2) above yields

$$D(F) = -\log Q(F) \qquad (4)$$

if, and only if, there are no inter-fragment associations. If

it is now assumed that the strongest association is that between F and its parent fragment, P, Hodes showed that the discrimination may be approximated by

$$D(F) = -Q(F)\log(Q(F)/Q(P)) \qquad (5)$$

if F is symmetrical, i. e. has only one possible parent fragment(180). In an operational implementation of this work(168), the rounded D(F) values were used to represent the number of bits to be assigned to each fragment in a superimposed bit screen system: thus fragments for which the ratio Q(F)/Q(P) was near unity were eliminated from the screen set since no bits would have been set. In the present work, which has been carried out for implementation in a dedicated bit screen system, such low discrimination fragments are automatically eliminated at screen set selection time; this is achieved by only allowing into the screen set those filial strings whose frequencies of occurrence are sufficiently differentiated from those of their parents.

IV.3 The description of chemical substructures by integer strings

The Sheffield group made a systematic investigation of the frequencies of a limited number of simple fragment types, these including simple, augmented and bonded pairs, octuplets, elements, coordinated, bonded and augmented atoms, linear four atom strings and simple ring descriptors(92,108,130,137,175). For each fragment type, frequency counts were made for all features at the lowest levels of description and then the most frequent fragments were investigated at the more detailed levels of specificity: thus a frequently occurring simple pair would be considered for inclusion at the augmented pair level(179). The hierarchical nature of the fragment types was thus reflected in the method of screen selection which was performed manually using the ranked fragment frequency lists.

The fragments considered at WRAIR covered a much wider range of substructural sizes and the hierarchy was much less well defined though no mention is given as to whether this affected the ease of query encoding. The initial fragments were individual atoms and single adjacent atoms were added to these to form the fragments in the first iteration: thus a tetravalent atom would give rise to four filial fragments. However, presumably because of the vast number of fragment types produced, increasingly severe restrictions were made as the number of iterations increased; thus after the second iteration, the fragments were limited to unbranched acyclic chains and monosubstituted rings. During each iteration, fragment incidences were cumulated and those occurring in less than 0.1% of the file deleted from furthur consideration; conversely, those occurring in more than 1% of the file were included in the subsequent iteration. This approach is very similar in concept to the methods developed by Salton and his co-workers for the automatic generation of indexing terms for

document description(182,183). It is not obvious from their report (168) how much manual work was required by the WRAIR workers but it seems clear that the entire file was processed at each iteration of the fragment generation algorithm.

The work described here represents a method for screen set generation in which not only are all the steps algorithmic in character but also it is possible to produce a screen set from a single pass of the structure file without any subsequent manual intervention. The screens may be atom, bond or ring centred, are symmetrical and, within each fragment type, form a strong hierarchy to facilitate generic coding at search time. The process consists of three stages, these being the generation of all possible fragments at the most specific level of description, cumulation of the individual fragment occurrences to obtain frequencies covering the whole file and then selection of certain of these fragments for inclusion in the final screen set, this being carried out upon the basis of the incidence and association considerations outlined above.

It is clear that a large amount of sorting will be required to obtain the fragment frequencies for the screen set generation algorithm. The computational requirements are somewhat reduced in the present implementation since only the most specific fragments are generated, and hence need to be subsequently sorted, in the first stage of the procedure. The subsequent cumulation then considers not only the specific fragments actually present but also the more generic fragments from which they have been derived. Even so, the overall process will be most efficient if the fragment representatives are chosen to be as simple as possible to allow of rapid sorting prior to the second stage. Such an approach will also bear fruit in the initial fragment generation

step since it seems certain that simple fragment representatives will also be simple to produce. Whilst it is relatively easy to encode small fragments such as augmented atoms, a description of a large substructure, such as a non-generalised octuplet(137), requires, in effect, some form of connection table which raises two immediate problems. Firstly, the record must be converted into a canonical format with the minimum of effort: reviews of canonicalisation routines have been given by Bersohn(184) and Jochum and Gasteiger(185). More serious, in view of the very large numbers of records involved, is the sheer bulk of the record. Consider the substructure shown in Fig. IV.1: the circles represent substructures of increasing size, and hence fragments of increasing specificity, and may be considered as three levels of description for the double bond at the centre of the substructure. An explicit description of the most specific fragment would require a connection table involving 11 atoms and 10 bonds; still larger records are of course possible though their presence in the final screen set would be unlikely for all but the largest files. The main problem is hence one of compacting the representation whilst retaining easy access to the more generic, parent fragments contained within the substructure.

The fragment descriptors developed in the present work are strings of integers, each of which represents a more precise definition of the environment of the features described by the first integer in the string. The integers in the string are obtained by an adaption of the Morgan algorithm(141); as described in the third chapter of this thesis, the algorithm discriminates between atoms upon the basis of their extended connectivity values where the nth order connectivity of an atom is calculated by summing the (n-1)th order connectivity values of all adjacent atoms. Increased specificity is obtained if the (n-1)th value of

the central atom is also taken into consideration: thus, using the nomenclature of Chapter III, we may write

$$v_{a_i}^{n} = v_{a_i}^{n-1} + \sum v_{a_j}^{n-1}$$

where the summation is over all the atoms, a_j, which are adjacent to a_i. Consider the structure shown in Fig. IV.2 where the numbers attached to each atom represent the initial connectivity values. Two iterations of the algorithm yield the sets of values shown in Fig. IV.3. If we consider the oxygen atom, we may describe it by the string of integers (2,6,20); similarly the substituent methyl group carbon may be represented by (1,4,13). Note that these descriptions are purely topological and say nothing about the nature of the atoms that they represent or the order of the bond connections.

As noted in Chapter III, significantly increased discrimination between the atoms in a molecule may be obtained if properties additional to the connectivity are used to determine the initial property values. The initial values used in this work are numbers descriptive of the atom type and the number and types of adjacent bonds, i. e. the number describes a bonded atom; these integers are then used to provide higher order descriptions of the atoms. The structure representation used was the Crossbow connection table. The heart of this record is the units section which consists of a string of symbols, each of which is associated with one of the non-hydrogen atoms in the molecule, and the initial property values were simply the binary representations of these symbols. It should be noted that a very large amount of *a priori* selection has been carried out in the design of the units notation since frequently occurring atom types are assigned a variety of symbols to reflect the variety of bond surroundings that may need to be taken into account for adequate discrimination; rarer atom types or bond configurations are, on the other hand, generalised and only a

limited number of symbols are employed(186).

The Morgan algorithm can, of course, be iterated as many times as required so that it is necessary to define the maximum level of description that is required, i. e. the length of the integer string. The smaller, more generic fragments may then be obtained by successively replacing the righthand-most integer by zero. Initial experiments, using the basic approach described above, showed that refinememnts were required in that very few of the longer strings were found to occur more than one or two times. The circular substructures described by the non-zero part of an integer string representing an atom and its environment increases in size by one bond in radius for each iteration of the algorithm. Previous work, using the sample file of 30000 compounds mentioned earlier, has shown that the variety of atom-centred fragment types increased from 68 for atoms to 136 for co-ordinated atoms, 313 for bonded atoms and then suddenly to 2331 for augmented atoms (108). Many of the larger fragments were of very low occurrence: thus 960 of the augmented atoms occurred only once in the sample file. As the increase in variety from the first to the second level of fragment description in the present work may be expected to be of at least comparable suddeness, it is clear that very many of the larger substructures will occurr very infrequently. The problem was resolved by inserting two initial levels of description prior to the bonded atom representative, these two levels corresponding to the atom type and atom type plus connectivity. Thus the substructures described are in the regular progression shown in Fig. IV.4 with the first four levels representing elemental type, co-ordinated atom, bonded atom and augmented atom respectively; the version of the algorithm used to generate the fourth and subsequent property values was

$$V_{a_i}^{n} = 3*V_{a_i}^{n-3} + \sum V_{a_j}^{n-3}$$

However, only the integers from $V^3_{a_i}$ upwards, to a maximum of $V^7_{a_i}$, were actually written out to tape for subsequent sorting; thus the minimal level of description in the final screen set is the bonded atom.

So far, we have only described atom-centred fragments but the procedure is clearly applicable to any type of fragment given appropriate numerical substructural descriptors. Analogous bond-centred strings were produced from the atom-derived integers using the equation

$$V^n_{a_i a_j} = V^n_{a_i} * V^n_{a_j} \qquad (1 \leq n \leq 6)$$

where $V^n_{a_i a_j}$ is the nth order property value of the bond connecting a_i and a_j. Ideally, the first value should be the bond type itself(single, double, aromatic etc.) but this information may not be readily obtainable from a bond-implicit structure representation. With this proviso, the first four levels correspond to simple, augmented and bonded pairs and non-generalised octuplets, all of which have been used in previous work in Sheffield. It should be noted that a furthur type of bond-based fragment, the four atom string(137,187), cannot be produced using this method of fragment generation unless the corresponding substructure happens to be linear.

The great advantage of the technique over other methods of fragment generation is that no path tracing algorithms need to be invoked to detect the larger fragments since only the adjacent atoms need to be considered at each step: as noted by Barnard(188), this can make quite enormous reductions in computer time possible. Strings, analogous to those above, may be produced for rings but here, ways must be found to identify the monocycles that are present in the structure. As before(see pages 32-33), the subset of the rings described by WLN was used since these are

rapidly identifiable from the Crossbow record. The properties considered include ring size, number and type of heteroatom substituents, the number of extra-ring connections and whether the ring was fused: furthur details are given in the next chapter.

A final point that should be strongly emphasised. The integer strings developed here are structural descriptions which are intelligible only in machine terms: with the exception of the single integer strings, it is not possible to reconstruct the substructure corresponding to a given string. However, the set of strings has been constructed so that, hopefully, a range of highly discriminating screens, including both generic and specific descriptors, may be assigned to any input structure representation. It is thus ideally suited to systems involving direct structural input, such as by chemical typewriter(25) or an interactive graphics terminal(206), both for query encoding in substructure search and for the assignment of screens to a compound at registration.

IV.4 An algorithm for screen set generation

In the previous section, a method was outlined for the generation of substructural descriptors from a connection table representation of a chemical structure; we now describe how these descriptors may be used to produce an approximately equifrequently occurring set of screens for subsequent assignment and substructure search. The procedure is in three stages.

The first step is very simple, albeit the most time-consuming part of the procedure, and involves the analysis of the connection tables in the structure file. For each molecule used, integer strings are built up for all occurrences of the fragment type under consideration. Once this has been done, the strings are written out to tape for subsequent sorting upon an incidence basis, that is only a single occurrence of each string type is output per molecule since the subsequent screen assignment is to be upon a present/absent basis. It may be noted that, for all but the most common substructures in a compound file, the incidence and occurrence figures are not very different. Strings are only written out at the most detailed level of description that is required; the production of all the substrings as well results in a very large number of additional records for sorting.

Once all the strings have been generated they are sorted into increasing order of the integer strings so that all occurrences of a given string type appear together on the tape. These occurrences are then cumulated for each fragment and a simultaneous count is made of the less specific fragments which may be generated from the string, i. e., the string (6,1,100) will give rise to the substrings (6,1,0) and (6,0,0). The strings, together with their associated frequencies, are then written out to a second tape for sorting into ascending size and decreasing

incidence order where the three strings above are presumed to have sizes of three, two and one unit respectively. The resulting ranked frequency list is then used as the input to the screen set generation program which yields a set of approximately equifrequently occurring screens.

The derivation of sets of equally frequently occurring sets of attributes has been thoroughly investigated in the context of bibliographical information systems where the objects are textual in nature, e. g. document index terms or author names in a directory, and the attributes to be considered are strings of alphanumeric characters. Perhaps the most common approach is to represent the objects, i. e. the text, by variable length character strings, the longer strings representing those character juxtapositions that occur most frequently in the text corpus(189,190,191,192). In one application, character strings are generated from the text by moving along it one character at a time and producing a fixed-length string at each point; this length is the maximum-sized string that is to be allowed into the final set of attributes. The string occurrences are summed, together with those of their parent strings, and the resultant frequency list used as input to the symbol set generation algorithm. As text is one-dimensional in character, the attributes are also linear and generic attributes are easily obtained by successive righthand character truncation. Thus the first character of the word COMPUTE will yield the strings COMPUTE, COMPUT,,COMPU , COMP , COM , etc. down to C ; a detailed description of this procedure is given in (196).

The integer strings described above may be manipulated in an entirely analogous manner and we now present an algorithm, which has been implemented in Algol68-R, to produce a screen set from a sorted input tape file of string frequencies.

The program makes use of a threshold frequency, T, above which strings will be considered for inclusion in the screen set. The relationship

$$T = N/4*M$$

was used where M is the screen set size and N is the total number of fragment incidences summed in the cumulation program; a tape record containing this number is constructed so as to move to the top of the cumulated fragment frequency list after sorting and it will hence be the first record to be read by the program after the input tape file has been opened. The value of 4 in the denominator was found empirically: a similar relationship has been used by Yeates(197). The value of M is the only parameter required by the program and is usually one less than a multiple of 24; the computer used for this work had a 24-bit word-length and a single bit is reserved for use as a conflated screen, that is one which may be assigned if no match can be obtained for a substructure with any of the other screens in the set. An alternative procedure would be to ensure that all the single integer strings, i. e. all possible bonded atom representatives, were included in the final screen set.

A trial screen set is obtained by including in the set all those single integer strings with associated frequencies $\geq T$; as the number of such strings is usually less than M, the set is made up with dummy, zero-filled strings. During subsequent iterations of the algorithm, the set of strings of a given size is read in from the sorted tape file and these strings are considered for inclusion in the set so as to improve its equifrequency properties. Consider a general string, s_i, of length n whose parent fragment s, of length (n-1), is included in the set created at the end of the previous iteration. s_i

will be stored for consideration as a potential new screen if both its frequency, f_{s_i}, and the difference in frequency between it and its parent are not less than T, that is

$$f_{s_i} \geq T \quad \text{and} \quad f_s - f_{s_i} \geq T .$$

The latter requirement is to encompass the parent-filial associations discussed by Hodes(180) whilst the presence of the parent in the screen set is dictated by the need for a strict fragment hierarchy to permit easy generic coding at search time; it has also been claimed that, for character strings, an emphasis upon the shorter strings may yield a better final relative entropy(199). Since, in later iterations, there may be many possible strings satisfying the frequency criteria above, only the M most frequently occurring strings are actually stored; thus if m n-length strings have already been stored, the new string s_i will be discarded unless f_{s_i} is greater than the frequency of the least frequent of the strings already stored for subsequent consideration.

At the end of an iteration, that is at the end of the screens of a given size n, the potential new screens are merged with those already in the screen set and a pruning procedure carried out to remove certain superfluous screens. Consider an iteration in which strings of length n have been considered and then added to the (n-1) and smaller strings already in the set. Then for every (n-1)-length string, s, a check is made to see whether the addition of the n-length string s_i has reduced the parents' frequency below T, i. e. whether

$$f_s - \sum f_{s_i} < T$$

where the summation is over all the n-length filial strings, s_i, of the parent fragment. If this inequality is found to hold, filial fragments are deleted in inverse frequency order until

f_s rises above T. Thus if a string (23,179,60,0,0) has an associated frequency of 172 and its filial strings (23,179,60,473,0), (23,179,60,479,0),(23,179,60,515,0) and (23,179,60,720,0), with frequencies of 41, 74, 23 and 21 respectively, have been selected as possible new strings then

$$f_s - \Sigma f_{s_i} = 13$$

which, for a threshold of 20, is too low. The least frequent string, (23,179,60,720,0), is accordingly deleted and f_s now rises to 34; this above the threshold and the next (n-1)-length string may be considered after f_s has been reset to its original value.

After all the strings have been inspected in this manner, a note is made of the size of the current screen set and, if $\leq$ M, the set is sorted into alphabetical order, to permit a rapid lookup via a binary search in the next iteration, and the program continues to consider the (n+1)-length strings. If, however, the size of the new set $>$ M, the least frequent screens are deleted till the required size is achieved; the program then proceeds as before.

Occasionally, it may not prove possible to produce a screen set of the required size since there may be insufficient strings obeying the strict frequency requirements; in practice, this only appears to occur if a large set is being constructed from a very small compound file. Thus, Gannon found that a minimum of about 100 compounds was needed to produce sufficient six integer strings for a 239-member, atom-centred screen set(138).

Having described the procedure in qualitative terms, we now present the basic algorithm. Apart from s, s_i, f_s, f_{s_i}, T and M which have been introduced above, two arrays, A and B, need to be defined. A, which is of size M, is used to store potential screens of length n during an iteration whilst B, of size 2*M, holds the screen set obtained at the end of the (n-1)th iteration

together with excess space to accomodate the contents of A when the two arrays are mereged.

(i) read first tape record and M; calculate T.

(ii) n := 1.

(iii) read a string, s_i; if $f_{s_i} \geq T$ store s_i in B; repeat until all the single integer strings have been read and then sort B into alphabetical order.

(iv) n := n + 1; y := 0.

(v) read a string from the sorted input file; if the length $\neq$ n go to (ix)(dummy records have been inserted between each of the groups of strings of a given length).

(vi) if $f_{s_i} < T$ or s is not in B or $f_s - f_{s_i} < T$ go to (v).

(vii) if A is not full, add s_i to the (y+1)th position then if y = M, sort the strings in A in inverse frequency order and go to (v).

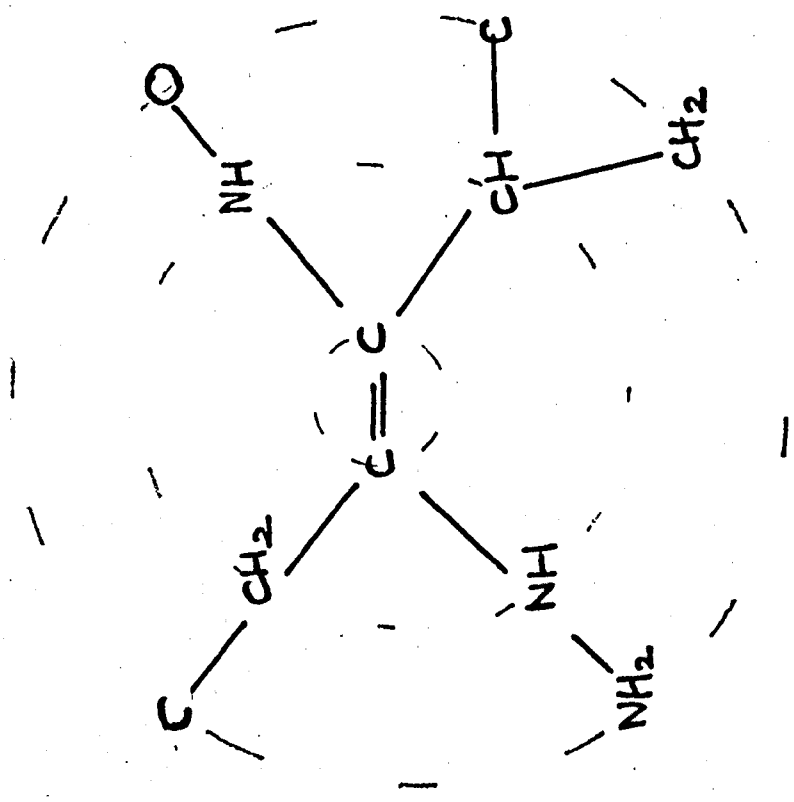
(viii) if $f_{s_i} >$ least frequent member of A, insert s_i and resort A; go to (v).

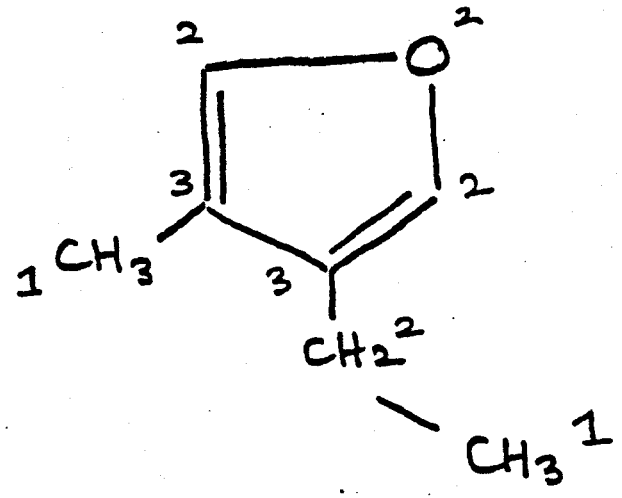
(ix) merge A with B and sort into alphabetical order so that each parent string, s, appears before its filial strings, s_i; for each string s, of length (n-1) evaluate $f_s - f_{s_i} \geq T$ and if the inequality does not hold, delete filial strings in inverse frequency order until true and then reset f_s to its original value.

(x) while the new set size $>$ M, delete strings, in size and inverse frequency order, from B; sort B into alphabetical order; if there are still strings to be considered go to (iv).

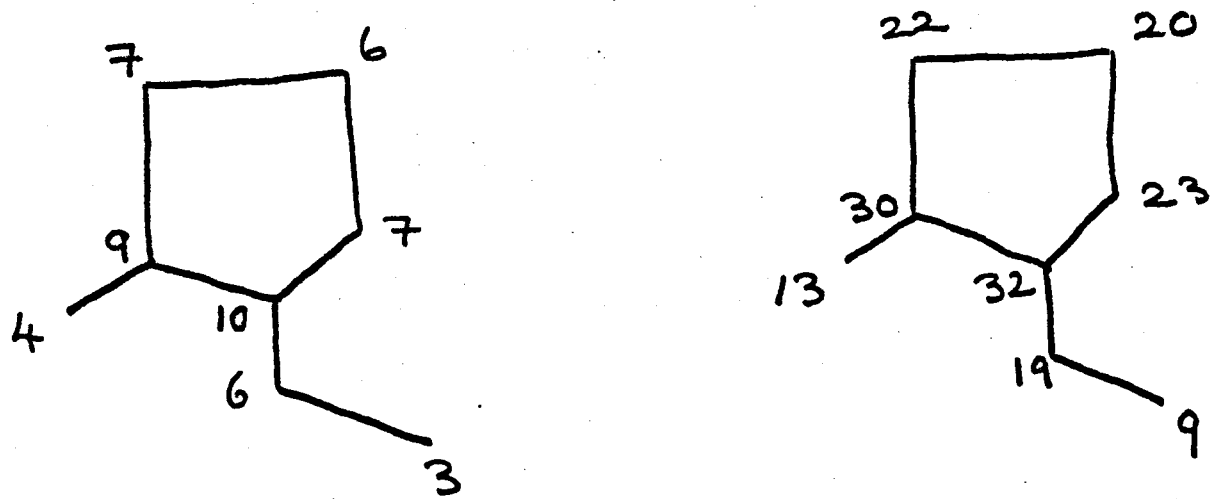
(xi) if the set size = M output B and halt; otherwise fault.

A listing of an Algol68-R implementation of this algorithm, using different notation, is included in Appendix I.

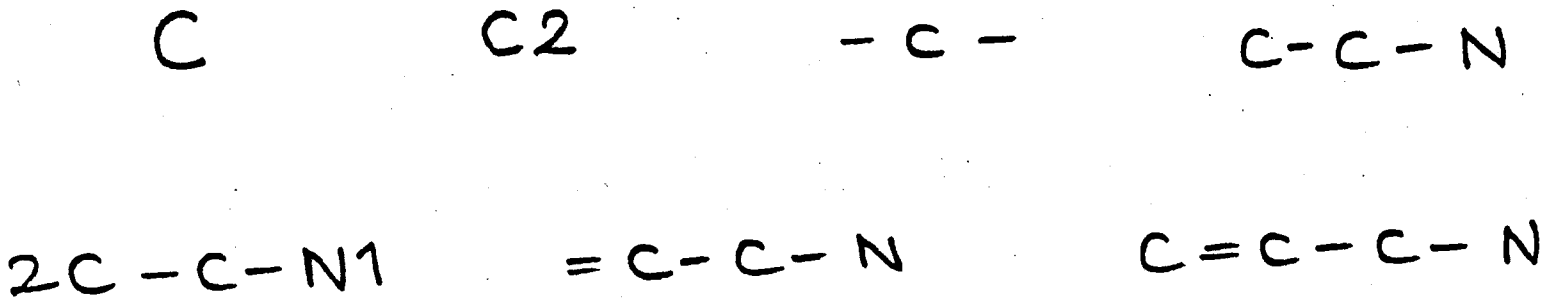

O
NH
C
CH
CH2
C
C
CH2
C
NH
NH2



First order connectivity values.



Second and third order connectivity values.



Typical substructures described by n-integer strings(n = 1 to 7).

CHAPTER V

A substructure search system for the retrieval of chemical reaction information.

V.1 Introduction

In the second and third chapters of this thesis, we have outlined two methods for the automatic characterisation of the structural changes occurring in a chemical reaction. In this final chapter, we describe the design and implementation of an experimental substructure search system(SSS) which permits searches for reaction queries to be made utilising both methods of analysis.

Many SSSs have been described in the literature(22,23,24,25,58, 181,193,194) but they all have in common the concept of two or more levels of structural description. This permits the efficient elimination of many obvious non-hits at an early stage in the search by the use of a simple binary attribute characterisation; subsequently, those molecules not so deleted are passed on for a more sophisticated, and time-consuming, analysis using a more detailed structure representation. The initial level of structural description is some form of fragment bitscreen which can be searched at very high speeds and the fragments described are generally algorithmically assigned from an input whole structure representation although manual assignment is, of course, also possible(125). In the previous chapter, we have discussed one approach to the design of screening systems for files of connection tables but others are, of course, possible(22,23,168) whilst Granito et al.(112) and Granito(113) have described an assignment procedure whereby the substructural features described by the Ringcode fragmentation method are specifically searched in a connection table; similar work has been reported elsewhere(24, 198). The bitscreens may also be automatically assigned from a linear notation(195,166,200). A recent review of twenty different screening systems is given by Powell et al.(201).

The second level of representation is often WLN which may be interrogated using conventional textual string search methods(103, 202). Computationally, this is currently a much slower procedure and, for certain types of substructure, a large number of alternative query character strings may need to be formulated for matching against the compounds on file. Granito et al.(166) quote a twenty-fold increase in processing time if an initial bitscreen search is not carried out and similar figures have been given by Sheng et al.(203).

Those structures that have satisfied the string search requirements may then be passed on for atom-by-atom matching; in many cases, the connection tables that need to be searched are produced in situ at this stage. Increasingly, however, the WLN level is omitted and the connection tables of those compounds satisfying the initial bitscreen requirements are matched directly against the queries(22,25,204,205): in such cases, a greater amount of effort must be expended to ensure an adequate choice of fragments in the bitscreen. Indeed, Rowland and Veal(59) have recently shown that acceptable retrieval results may be obtained simply at the bitscreen level: experiments with a sample file of over 200000 compounds showed that a screening system based on the principles discussed in the previous chapter achieved an average precision of over 60% for a wide range of substructural queries.

The reactions SSS to be described is of this latter form, the heart of the search file consisting of a series of bitscreen records, though limited WLN stringsearch facilities are also provided. It should be noted that the search file is small, less than 4500 entries, in comparison with the compound files discussed above which, typically, have upwards of 50000 members. In the first chapter, we emphasised the multifarious nature of chemical

reaction data and the need to provide a variety of modes of access to it. Also important in a reactions retrieval system is the ability to differentiate between those substructural features that have been involved in the reaction and those that have not. Due to the use of fragmentation-based methods of reaction analysis, work in this department to date has only allowed access to the former type of moiety and this is also in large part true of the WLN approach of Chapter II. Given a computerised SSS, the unchanged features could, perhaps, be identified using the parent WLNs but in the printed format we have chosen to concentrate our attentions upon the features that have been involved in the change.

Our structure matching algorithm provides a unique opportunity to make explicit the differentiation between the reacting and non-reacting substructures since it identifies just those atoms involved in the reaction site without any simultaneous rupture of the reacting molecules. Given this ability, we may think in terms of a retrieval system in which we may carry out searches specifying either, or both, types of feature for the reactant and/or product structures. The SSS to be described below has been designed with this end in mind.

The next section, which is highly implementation-biased, outlines the way in which the various screen sets, which were to be used in the SSS, were generated; this is followed by a description of the query encoding routines and the generation of the search file and the final part of the chapter presents the results of applying a series of reaction queries to the search file.

V.2 Generation of screen sets

The source file for this work was those 7415 reactions successfully analysed by the WLN program described in Chapter II. For each such reaction, the WLNs of the reactant and product molecules were input to a Crossbow connection table generation program. If tables were produced for both of the molecules involved in the reaction, the WLNs, connection tables and WLN analysis fragments were written out to tape; in this way a file of 5226 reactions was obtained in Crossbow table format for subsequent processing.

An analysis by Clews(49) showed that, as might be expected, a file of reaction site residues contained a higher percentage of heteroatoms than the corresponding file of parent compounds. It was hence decided to use a different set of screens for characterising the atoms and bonds in the reaction sites from the sets used for the atoms and bonds in the parent molecules(a subsequent comparison of the atom and bond centred screen sets in each case showed quite marked differences: thus of 239 screens in each set, 38 atom and 50 bond screens were found not to be common to the two sets).

We have selected four different types of screen for assignment for both the parent molecules and for the reaction sites; the screens are based upon atoms, bonds, rings and molecular formulae. As these four types of screen are to be applied to both reaction site and molecular features in both the reactants and the products, a total of 16 different screen sets are ideally required. However, inspection of the index entry fragments arising from the WLN analysis showed little difference between the reactant and product fragment frequencies of occurrence; it was hence decided in each case to use the same screen set for assignment to features on both sides of the equation. There are thus a total of eight screen

sets to be assigned to each reaction: this is almost certainly excessive in view of the small size of the file but this is preferable to underscreening which seems to have been a severe problem in a previous reactions SSS developed in Sheffield (100).

The two types of molecular formula screen are easily obtained from the numbers and types of atoms in the parent molecules and the reaction sites. The screen is a single, 24-bit word record and contains entries for carbon, nitrogen, oxygen, sulphur, phosphorus and the halogens together with bits for general halogens and general heteroatoms. It should be noted that these screens are based primarily upon chemical intuition rather than the methodology of Chapter IV; thus screens are available in the reaction site screen for the presence of upto 6 carbon atoms whilst the sulphur screens are restricted to 1 or more than 1.

The ring screens have been obtained from two different sources. The Crossbow record permits the easy identification of those rings present in a molecule which have been explicitly delineated by WLN and we have used these descriptions as the basis for the molecular ring screens. A maximum of three levels of description has been used, the first and second of these being the ring size and the number of heteroatoms. If the ring is not carbocyclic, the third level describes the atom types present in the ring but if there are no heteroatoms, the level decribes the number of extra-ring connections. To explain why this differentiation should be made, consider a carbocyclic, six ring; the first two integers in the string will be 6 and 0, and nothing that we put in the third position can add any further information about the ring characteristics if the integer pertains to heteroatomic data and a different type of data is hence encoded. The heteroatoms

nitrogen, oxygen, sulfur and phosphorus are assigned the arbitrary values of 1000, 100, 10 and 1 respectively and the third level description is obtained by summing the values for all such atoms present in the ring. Thus the two monocycles shown in Fig. V.1 will be described by the strings (3,1,100) and (6,0,3) whilst the ring system will yield the strings (6,0,2) and (5,1,10). As an implementation detail we should note that the second integer in the string has 50 added to it. Consider a ring giving the string (6,0,3): in the course of the screen set generation program this will be decomposed to the one and two integer strings (6,0,0) and (6,0,0) which cannot be differentiated. Rewriting the originalstring as (6,50,3) gives rise to (6,50,0), a carbocyclic six ring, and (6,0,0) which is simply a six ring.

It would be possible to obtain analysis ring screens, that is descriptors of the rings that have been changed in the course of the reaction, by a comparison of the rings of the reacting molecules using the Crossbow record in much the same way as we have for the derivation of molecular ring descriptors. This has, however, been carried out already in the course of the WLN analysis and the non-common rings are included in the analysis fragments for each reaction; moreover, these fragments include the synthetically important ring carbonyl function(though the presence of this feature could easily be deduced from the Crossbow record if the WLN analysis had not already been carried out). As before, three levels of description are used in the construction of the integer strings, the first two being the ring size and the number of heteroatoms. For carbocycles, the third integer represents the saturation which is described by the presence, or otherwise, of a T symbol immediately prior to the J ring delimiter in the fragment WLN symbol string; if the ring is saturated,

the third integer is 1, otherwise 2. For heteroatomic rings, account must be taken of the following seven WLN symbols: V, O, S, N, M, K and P. These are assigned the arbitrary values 1, 10, 100, 1000, 1000, 1000(since all three symbols describe nitrogen) and 10000 and the second level integer is obtained, as before, by summation. For both hetero and carbocyclic rings one million is added to the third integer if the ring is fused, i. e. has @ as the first symbol in the fragment character string. If the ring is carbocyclic, the second integer is set to 99 for the reasons given above. Thus the WLN analysis fragments @T6 AV DVTJ, @L6J and T6 AM BSJ will be described by the strings (6,2,1000002), (6,99,1000002) and (6,2,1100)(though the actual screens assigned will, of course, depend upon the actual strings that have been chosen for inclusion in the screen set).

The actual analysis ring screen set was obtained from a list of the fragments obtained from the WLN analysis; it should be noted that the resulting screen set is probably different from the one that would have been obtained from the actual file of 5226 reactions used in this work since many of the WLN-analysed reactions were not processed successfully by the Crossbow program. Each ring in the list of analysis fragments, i. e. each string commencing with T, L or @, was processed symbol by symbol, to produce a three integer string as above and this was then written out to tape for string cumulation and, subsequently, screen set generation. The chosen screen set size was 46, i. e. two bits less than two computer words. One of these bits was used as a conflated screen whilst the other was used to describe the phenyl ring, the WLN symbol R, since this is not amenable to the integer characterisation described earlier.

The molecular ring screen set, on the other hand, was made from

a list of all of the different WLNs in the file of 5226 reactions used for this work. After the elimination of duplicate WLNs, a total of 8939 Crossbow tables were produced and these were processed to produce the appropriate three integer strings. Also, after conversion to redundant adjacency matrices using a program written by the author, these tables were employed to produce the integer strings which formed the basis for the molecular atom and bond screen sets. The preparation of these has been described in Chapter IV; suffice it to state here that the maximum length strings considered were six integers for bond-centred fragments and five integers for atom-centred ones which correspond to the substructures illustrated in Fig. V.2.

For each molecule on the tape of 8939 compounds, all possible ring, bond and atom integer strings were generated at the highest level of specificity. For each string of a given fragment type, a check was made to ensure that an identical string had not already been written out to tape for that compound; this was done to ensure that the cumulated fragment frequencies referred to string incidences since the resultant screen sets were all to be assigned on an incidence basis. The connection table analysis program was written in Algol68-R and the single pass of the 8078 compounds on the tape for which redundant adjacency matrices could be produced required 1267 seconds of cpu time, this including the adjacency matrix generation and all the read and write accesses to four magnetic tapes. The resulting integer strings were subsequently processed to produce screen sets of 239, 239 and 47 members for the bond, atom and ring features respectively: in each case, the missing bit was used as a conflated screen.

The final screen sets to be discussed are those used to characterise the atoms and bonds in the reaction sites. For this

purpose, a one in three sample of the complete data base was isolated. For each reaction in this subfile, redundant adjacency matrices were produced for the reacting molecules and these matrices were then compared using the graph matching algorithm of Chapter III. Integer strings were then created for all of the atoms and bonds in the reacting molecules but only those relating to substructures entirely contained within the derived reaction site were written out to tape for subsequent screen set generation. Thus for the reaction shown in Fig. V.3, only the atoms and bonds in the partial structures shown in the lower half of the Fig. were considered. Note that although the central feature, atom or bond, is contained within the site, the integer strings may well, in the later integers in the string, describe features outside of the site; thus a certain amount of environmental information is automatically encoded.

As described in Chapter IV, one measure of the effectiveness of a screen set is the relative entropy, this being a measure of the equifrequency of the screens in the set. However, the sets were obtained from files different to that to which they are to be assigned, i. e. the search file, and thus the relative entropies quoted in Fig. V.4, which are based upon assignment to the appropriate source file, would probably be somewhat different if search file assignments were considered. The source file dependence is made manifest by the relative entropy of the analysis ring screen set which is markedly different from the corresponding molecular ring screen set. The difference is presumably due to the fact that the source file for the analysis ring screen set, a list of the WLN analysis fragments, did not contain frequencies of occurrence, merely the presence of that fragment somewhere in the file of 7415 reactions processed.

The entropies in Fig. V.4 are based upon non-redundant assignment, that is if a screen (6,50,2) is assigned then the parent screens (6,50,0) and (6,0,0) will not also be set. For a search system, however, this additional coding must be done to permit easy generic searching and it is hence of some interest to see the extent to which the equifrequency performance is degraded as a result of these additional assignments. In fact, assignment of all possible screens to the source file for the three molecular screen sets produced relative entropies of 0.929, 0.786 and 0.923 for the bond, ring and atom screen sets respectively; this entropy reduction in all three cases is noticeably less than the increase when compared with the initial single integer strings.

V.3 Creation of the search file and query encoding techniques

A program has been written to process the file of 5226 reactions mentioned above in section 2. It is in three main parts. The first segment takes a Crossbow connection table and converts it to a redundant adjacency matrix, at the same time assigning molecular ring screens to the compound using data in the Crossbow record; the process is repeated for both of the reacting molecules. This routine is incapable of handling Crossbow tables derived from spiro or bridged ring system-based compounds and is thus able to process only about 90% of the reactions in the file. In the second segment, the two matrices are compared using our structure matching algorithm as described in Chapter III. The final section validates the analysis, assigns molecular atom and bond and analysis atom, bond and ring screens to both molecules and then writes the bitscreens, WLNs, WLN analysis fragments and bibliographical details out to tape.

The program contains about 900 lines of Algol68-R code and has been run in 125K words of core on the University of Sheffield ICL 1906S computer under GEORGE 4. The program required 2428 seconds of cpu time to process the file of 5226 reactions, that is just over two reactions per second, and the results of this computer run are given in Fig. V.5. These figures are very similar to those obtained from the sample file, and shown in Fig. III.10, in that circa 93% of the reactions on file are processed by the structure matching algorithm; by extrapolation from the sample file it may be expected that about 45 of the analyses are in error in some way. The subtotal of 37 detected failures in the Figure corresponds to those reactions in which all of the reactant or product atoms were eliminated in the course of the matching procedure: such an occurrence should not take place if

valid mappings have been identified(see, e. g., the reaction of Fig. III.20).

The time required, about half a second for each reaction, is about seven times that needed for the corresponding WLN analysis. However, the two timing figures are not directly comparable for at least three reasons:

(i) real magnetic tapes were used as against the GEORGE filestore multifiles(207) which were employed for storing the results and the data for the WLN program; the latter procedure is very much faster.

(ii) the Algol68-R compiler used here produces object code which is about two thirds as fast as that produced from a comparable COBOL source program.

(iii) the program needs to be used with full run-time overflow checks to encompass those reactions in which the atomic property values become too large for the computer word reserved for them.

It should also be noted that the actual degree of program complexity was considerably less in the earlier program, the majority of which consisted of MOVEs and comparisons of various types. An annotated listing of the Algol68-R program is included in Appendix II.

The file of 4388 reaction analyses was then used as the source file for the reactions substructure search system.

The software that has been developed to search the file of reactions is, of course, in large part identical to that used to characterise the analyses. In particular, bitscreens are assigned using a connection table as the primary input query structure representation. These tables are processed in just the same way as the adjacency matrices of the reacting molecules to produce integer strings which may then be matched against the appropriate screen set(s). If a match is not obtained for a search string at the maximum level of description, the

query representative is shortened by one integer using righthand-most truncation and the assignment procedure, a binary search routine, called again. The process continues until a matching string is found, when the appropriate bits are set in the query bitstring, or the conflated screen is assigned.

The input connection table must be complete, i. e. not have any unspecified connections. As the majority of queries involve substructural features, means must be found to satisfy the unspecified attachments. Since the atom, and hence the bond, screens are based upon the units values of the atoms under consideration, such unsatisfied valencies in the query may be filled by the use of a dummy atom with the units value ?, a symbol not used in the Crossbow system. This being so, not only will no match be obtained if we search such an atom against a screen set but also we will not assign screens corresponding to substructural features larger than than that explicitly delineated by the query. To explain this point, consider the query substructure shown in Fig. V.6 in which ? represents an unspecified atom. If we consider the carbonyl oxygen atom, we wish to assign screens corresponding to the circular substructure shown at (a) in the Fig.; and no larger. Such a substructure will be described by the first four integers in the atomic property string, the remaining one describing a substructure that contains the dummy atom. Since the exact demarcation point at which the required feature ends can only be determined by some form of path tracing algorithm, the full, five integer string must be generated and searched against the appropriate screen set; however, the presence of the ? units value causes a contribution to the oxygen atom's fifth order property value which ensures that a match will not be obtained with the screen set at this level of substructural description. Similarly, no match can possibly be made for the acid carbon atom

except at the first level due to the adjacent dummy atom. Entirely analogous arguments apply to the assignment of bond centred screens since these are derived from the atom centred property values.

As well as atom, bond, ring and molecular formula searches, provision is also made for WLN string searches of the analysis fragments; it would also be possible to include string search facilities for the parent WLNs but this has not yet been implemented. String searching is only carried out after an initial bitscreen match to cut down on search times and these are furthur reduced by the use of a minimal requirements statement which is set at the front of each analysis and query. This statement is a single 24-bit word and gives details of the number of atoms and rings in the reaction sites and the number of rings in the reacting molecules, thus permitting the rapid elimination of many certain non-hits prior to the bitstring and WLN matching.

The search program is based on a simpleserial search technique (1) in which the entire file is matched against the queries, one record at a time. The set of query statements is held in core together with a series of associated binary hit vectors, one of which is assigned to each of the queries; each bit on such a vector corresponds to one of the reactions in the file. Each reaction which matches a particular statement has the appropriate bit set in the vector corresponding to that query statement. After the whole file has been traversed in this way, the logical operations are read in and the hit vectors merged accordingly using AND, OR or NOT logic. The tape containing the search file is then rewound and searched against the hit vectors: if a reaction is read whose bit has been set in one of the vectors, the WLNs of the reacting molecules and the bibliographical reference are output to the lineprinter together with the

number of the query that has retrieved the reaction. For many of the queries searched, the logical operations required were minimal and thus the second pass of the tape somewhat redundant; the adopted approach does, however, mean that the logical operations need to be considered only once, rather than after every reaction in the file, thus saving upon computer usage if many logical manipulations need to be performed. Search times were generally about 20 seconds for up to a dozen query statements, these including a variety of bitstring and WLN requirements.

The query encoding routines will be described primarily by examples, the first such query being shown in Fig. V.7. The query will be coded in two parts, corresponding to the two possible reactant reaction sites, and then the potential hits will be OR'd together. In both the cases, all of the atoms shown may be expected to be included in the reaction site and hence we need to search only the reactant and product reaction sites without any simultaneous inspection of the molecular bitstrings. An arbitrary numbering of the first reactant reaction site is shown in the lower half of the Fig. together with the accompanying connection table that is input to the search program. The first line, C 2 5, states that a connection table is to be input(C), that the reactant reaction site screen set and bitstrings are to be used(2) and that there are five atoms in the connection table. After this initial card, each subsequent line in the Fig. corresponds to an individual atom, a_i, in the query substructure; the first four integers give the numbers of the atoms that are attached to a_i and the following symbol, ?, Y, Q, C or Z in this case, is the units value of a_i. There follows a space, the connectivity of a_i, which is compared with the number of attached atoms as a check on the coding, and a true or false value depending

upon whether a_i is, or is not, in a ring of some sort. Analogous connection tables for the alternative reactant and the product reaction sites are produced in a similar way.

Although such connection table descriptions are complete in themselves, these representations may be made more precise by the inclusion of WLN strings in the query. These symbol strings may then be matched against the WLN analysis fragments of any reaction which satisfies the bitstring requirements. In the present case various WLN symbol strings could be used but, upon the basis of the ranked fragment frequency lists mentioned in Chapter II section 4, the following changes are used:

$$/\text{CN} \longrightarrow /1\text{Z}$$

$$\text{and } /1\text{NW} \longrightarrow /1\text{Z}.$$

The complete deck of cards used to search for this query are shown in Fig. V.8, the first card being a count of the total number of query statements that follow. Each such statement is preceded by a MIN card which contains certain minimal requirements that must be met before a bitstring and WLN search can be carried out; the six numbers refer to the numbers of reactant and product molecular and analysis rings and to the numbers of atoms in the reactant and product reaction sites. This record is coded in a single word at the start of each query and file record bitstring and is of most value where ring changes or large numbers of atoms are involved in the reaction; in the present case, few reactions will be eliminated from subsequent search. The W 2 and W 4 cards describe reactant , the 2, and product, 4, WLN fragment searches; to date, no facilities are provided for searching the reactant, 1, and product, 3, WLN strings.

Each query statement is terminated by a semicolon and the deck completed by the logical requirements that must be performed upon the individual hit vectors; the initial LOGIC card gives the number of sets of operations that are to be carried out. For the first, and in this case only, such card, the string of integers 1 1 20 states that query 1 has 1 logically related query, that this is query 2 and that the relationship is OR, i. e. the hits from the two queries are to be merged before output. AND and NOT logic are also available.

The second reaction is shown in Fig. V.9 together with the query coding used. In this case, the connection tables input, C 1 5 and C 3 5, refer to the molecular reactant and product substructures respectively and correspond to the requirement for an unchanged ring carboxyl function. The R 2 cards denote that a reactant WLN analysis ring string follows and this is processed to produce a three integer string as described above; also, the W in column 21 of the card specifies that a subsequent WLN string search should be made to ensure an exact match. The two ring strings obtained will be (6,1,1) and (6,1,1000001) for the reactant and product rings respectively; however the analysis ring screen set will describe both by the bit corresponding to (6,1,0), i. e. any monoheteroatomic six ring, and thus the WLN search will ensure a more precise query formulation. The MIN card should prove much more effective in this case due to the various ring requirements. The logic is carried out in two stages; firstly, the two possible rings are ORed together and then the resultant hit vector intersected with the hits resulting from the connection table search.

In many reactions involving ring changes, specification of these alone may often be sufficient, especially if a WLN search

is also possible. An example of such a query is shown in Fig. V.10.

Apart from connection table, ring and WLN string based queries, molecular formula requirements may also be input. These are normally calculated from any input connection tables but they may also be stated explicitly if a table cannot be used, e.g. if there are a variety of possible units values for one of the atoms in the table. Also provided is a U, that is units, facility which is employed if a single units type, rather than an entire connection table, is to be specified. The search program sets up the appropriate atom integer string and then zero fills all but the lefthand-most element; the string is then searched against a screen set in the normal way. Both types of query encoding are illustrated by Fig. V.11 where the prime requirement is the change of one oxygen atom to a nitrogen or sulphur atom and this is encompassed by the first two card sets. An M card contains a number, 1-4 as before, specifying the type of screen that is to be set and then nine integers which give the minimum numbers of carbon, bromine, fluorine, chlorine, iodine, nitrogen, oxygen, sulphur and phosphorus atoms respectively that are required. In this case, the two molecular formula change statements shown would be ORed together and then ANDed with the requirement for a reactant analysis, oxygen atom contained within a ring. This is specified by the format U a bcd where U denotes a units card, a the screen type(1-4), b the number of values which follow and then b pairs of characters,c and d, the first being the units value and the second, T or F, depending whether the atom is, or is not, contained in a ring. In the present case, our requirement would hence be described by U 2 1QT. The query may be completed, although this is not shown in the Fig., by a series of U cards for all of the possible

units values of nitrogen and sulphur atoms in a ring; these are ORed together and then ANDed with the hit vector resulting from the first three statements. In fact, this query retrieved six reactions, four of which were relevant to the query.

V.4 Evaluation

In this section we discuss the evaluation of the reactions SSS in terms of a set of 102 queries which were searched against the file of 4388 reaction analyses described in the previous section.

Three main sources were used for the compilation of the set of queries. The first of these was 34 real queries supplied by the Research Information Department of Pfizer(UK) Ltd.; an additional 3 queries were taken from Campbell(100) who states that they had been provided by chemists from ICI Ltd.(Pharamceuticals Division). The second group was the 18 queries used in the Derwent - WLN comparison of Appendix III. The remaining 47 questions were culled from a variety of literature sources, all of which contain illustrative reaction types that should be easily searchable in a reactions documentation service(26,27, 52,76,208).

The queries were coded up as described in Section 3 and then batched up, four or five at a time, for searching against the file of analyses; run times were typically 18 seconds inclusive of tape transput although one run, which contained a total of 27 different WLN fragments for stringsearch, required over 35 seconds.

In toto, of the 102 queries searched, 75 produced no output at all, whilst the remainder gave rise to a total of 643 retrievals, individual queries producing between 1 and 113 reactions. We have evaluated these results in terms of precision and screenout. Precision is widely used in document retrieval experiments(209,210, 211), normally in conjunction with recall, but screenout is a much more useful parameter for chemical retrieval experiments since the relationship between the form and the content of the machine-readable representation ensures perfect recall. For a file of

N records, n of which are retrieved in response to a query, we define the screenout, S, by

$$S = 100(N-n)/N.$$

Thus S represents the ability of the system to reject definite non-hits if the query encoding has been carefully carried out. Using this definition, it may be seen that the lowest screenout obtained, i. e. the largest number of retrievals(113), was 97.4% with all but 6 of the queries having screenouts or 99% or greater.

The determination of precision figures is somewhat more difficult if performed in the absence of user relevance judgements. In principle, an assessment of relevance may be made by an atom-by-atom search for an exact match with the query substructure but the figures obtained may be misleadingly high if other, concurrent changes have taken place in the course of the reaction(as defined by the set of reactant and product structures). We have hence considered as hits only those retrievals which exactly match the query statement and in which no other changes have taken place. Thus the reactions shown in Fig. V.12 will both be considered as false drops for the query shown in the upper part of the Figure.

With the proviso that the quoted precision values represent a lower limit to that which might be expected in practice, the results of the searches for those 27 queries which retrieved some material are shown in Fig. V.13. In this Figure, we show the query number together with four numbers, these being the number of reactions retrieved, the number of hits, the screenout and the precision, the last two being expressed as percentages and rounded to one decimal place for the screenout values. Although the figures for queries retrieving only one or two reactions are somewhat misleading, it can be seen that the system effectiveness is quite high with 17 of the queries

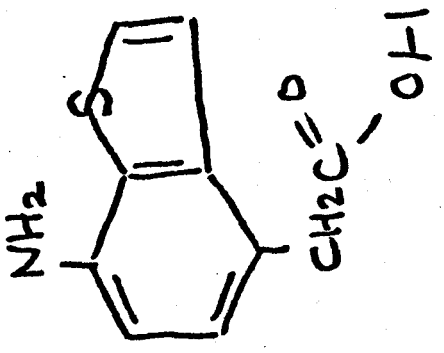
producing precision figures of 50% or more; 336 of the 643 retrievals were considered as hits, an overall precision of 52%.

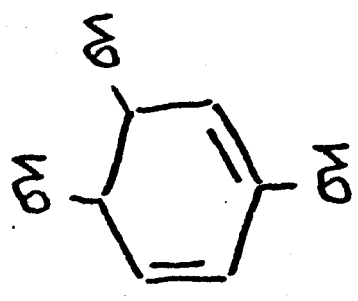
Only two of the queries, nos. 19 and 24, resulted in noticeably poor retrieval; the query requirements which produced 20 false drops out of 20 and 64 false drops out of 69, are shown in Fig. V.14 and it will be seen that both correspond to very general enquiries for which a minimal amount of query encoding is possible. Such reactions could be more precisely searched if an exact molecular formula change could be specified since at present, the two reactions can be searched only by stating that there are at least two carbon and one nitrogen atoms difference respectively between the reactant and product molecular formulae. Thus all reactions in which these minimal requirements are satisfied will be retrieved irrespective of the other molecular formulae changes: at present, this problem can only be overcome by a series of NOT cards to remove all other possible molecular changes.

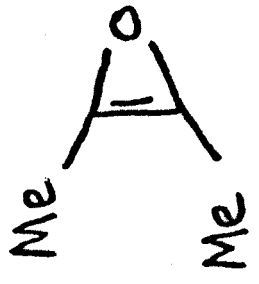
The distribution of retrieval set sizes is similar to that given by Adamson et al.(212), the majority of the queries producing little or no material at all. Such queries were often very specific in character, requiring the formation of specific rings or the reaction of quite complex functionalities; most of the queries in this class were from the group of real industrial questions. Conversely, a few of the queries produced a large output both in terms of actual numbers of reactions retrieved and in terms of the number of hits. Such reactions are generally quite simple in character as noted in the WLN analyses(see Fig. II.41) and by earlier workers(139,140); examples of such queries are shown in Fig. V.15.

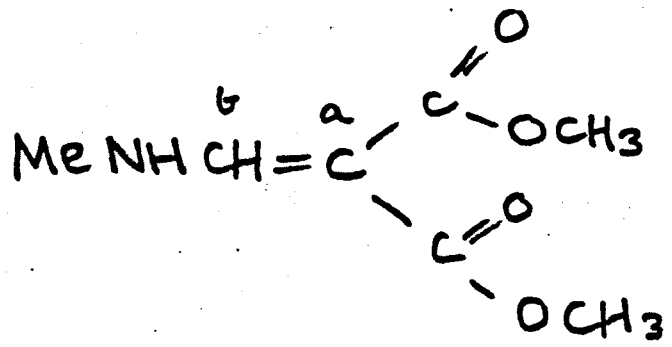
V.5 Conclusions

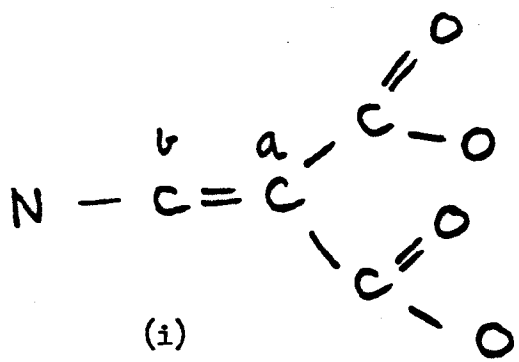
The high average screenout and precision figures obtained for the set of queries indicates the effectiveness of the screening system in providing rapid and accurate access to the data for a very large fraction of the substructural reaction queries that might be expected in an operational environment. The results may also be taken to show the ability of the structure matching algorithm to characterise the reaction sites within the pairs of molecules involved in a reaction. Moreover, the entire process of reaction site detection, screen set generation, screen assignment and search is fully automatic with manual intervention required only at the query coding stage. Even here, significant savings of user effort have been achieved by the use of a connection table as the primary input query medium. The tables are subsequently processed to produce atom, bond and molecular formula screens without the need for the coder to have any idea as to the contents of the various screen sets. In this respect, the system is similar to that developed by Feldman(149); indeed, the search mechanisms are ideally suited to online usage via an interactive graphics terminal. Useful additions to the screening system would be an exact molecular formula change facility,as described earlier, and a stringsearch capability for parent molecule WLNs; this latter utility would be primarily useful for the rapid detection of steroids, penicillins and other characteristic ring nuclei.


S
NH2
CH2C=O
OH

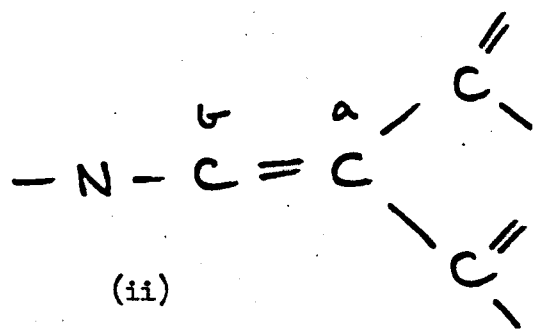

OH
OH
OH


O
Me
Me



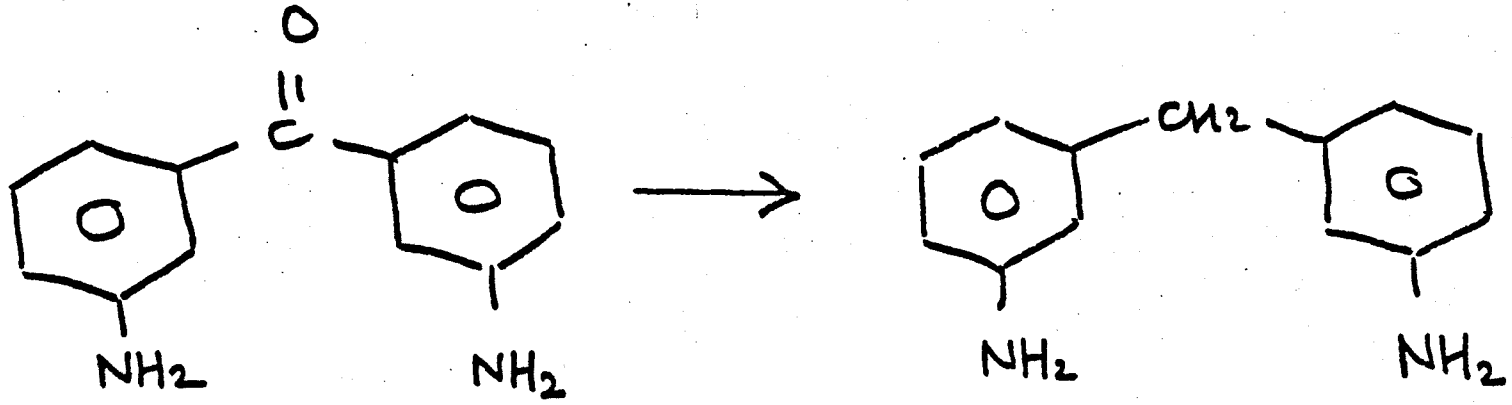


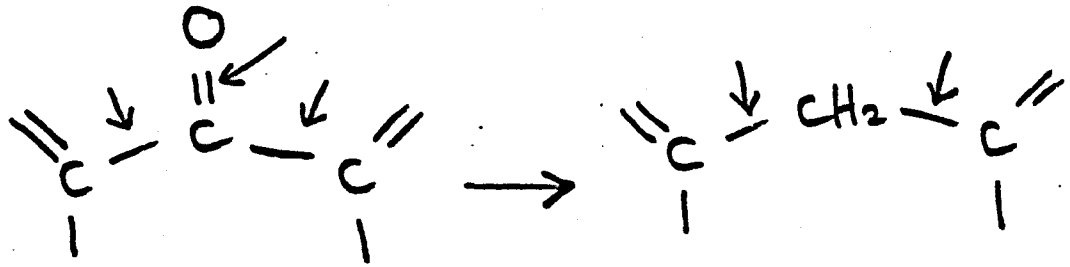
(i)



(ii)

Substructures described by: (i) a 7 integer atom string centred upon atom a(of which only the final 5 are considered - see text) and (ii) a 6 integer bond string centred upon a and b.





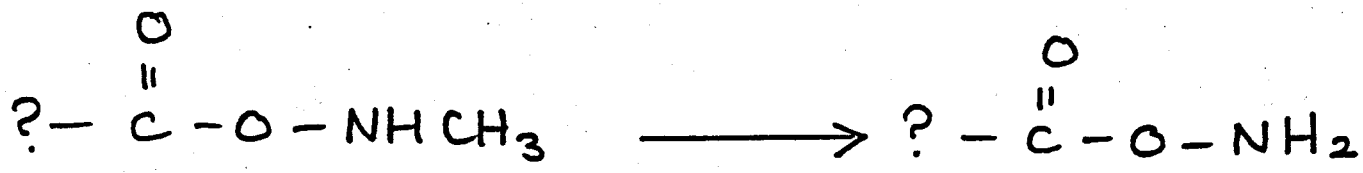
Only the bonds indicated will be considered for analysis bond screen set generation.

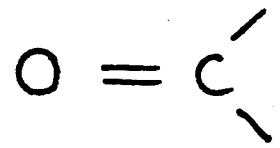
		Number of strings used	Screen set size	Relative entropy of single integers	Relative entropy of screen set
	Bond	171804	240	0.603	0.967
Molecular	Ring	19584	48	0.420	0.802
	Atom	160294	240	0.798	0.953
	Bond	16127	240	0.708	0.977
Analysis	Ring	1052	48	0.703	0.942
	Atom	21044	240	0.822	0.956

Generation statistics for the screen sets used in the reactions SSS.

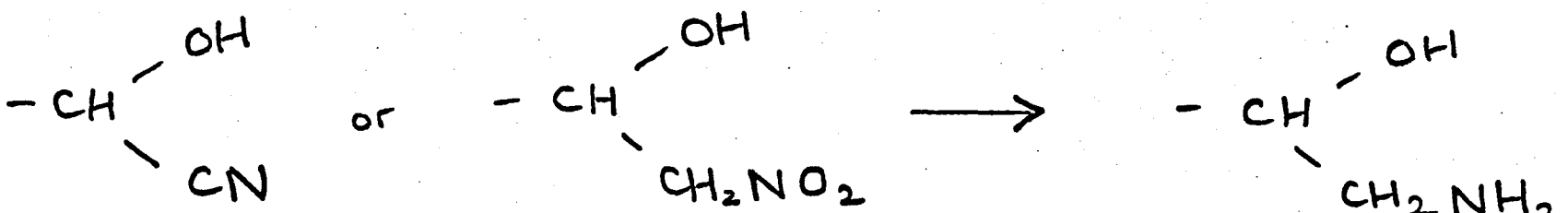
Reactions processed	5226
Adjacency matrices generated	4729
Successful analyses	4388
Overflow	8
No atoms matched	296
Detected failures	37

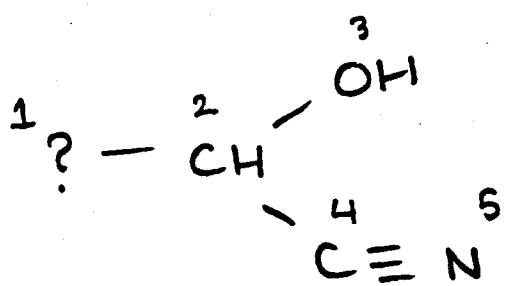
Creation of the search file for the reactions substructure search system.





(a)



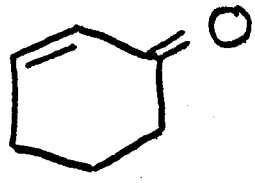


Typical query substructure and corresponding connection table

```
C 2 5
2 0 0 0? 1F
1 3 4 0Y 3F
2 0 0 0Q 1F
2 5 0 0C 2F
4 0 0 0Z 1F
```

```
2
MIN 0 0 0 0 4 4
C 2 5
2 0 0 0? 1F
1 3 4 0Y 3F
2 0 0 0Q 1F
2 5 0 0C 2F
4 0 0 0Z 1F
C 4 5
2 0 0 0? 1F
1 3 4 0Y 3F
2 0 0 0Q 1F
2 5 0 0L 2F
4 0 0 0M 1F
W 2
/CN                    /
W 4
/1Z                    /
;
MIN 0 0 0 0 6 4
C 2 7
2 0 0 0? 1F
1 3 4 0Y 3F
2 0 0 0Q 1F
2 5 0 0L 2F
4 6 7 0@ 3F
5 0 0 00 1F
5 0 0 00 1F
C 4 5
2 0 0 0? 1F
1 3 4 0Y 3F
2 0 0 0Q 1F
2 5 0 0L 2F
4 0 0 0M 1F
W 2
/1NW                   /
W 4
/1Z                    /
;
LOGIC 1
1 1 20
```

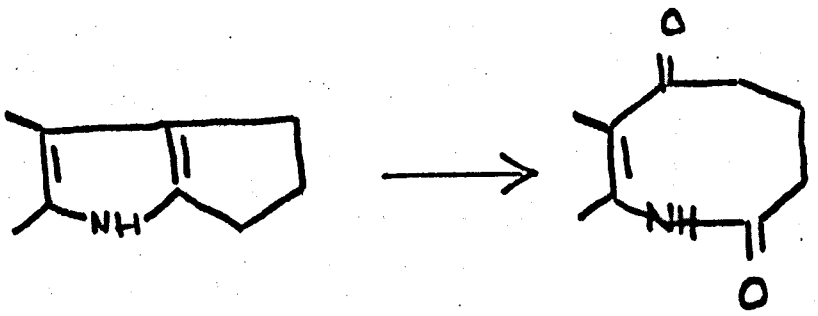
Complete query deck for the substructural change shown in Fig. V.7.



opened whilst a ring containing a carboxyl group, -CO.O-, is unchanged.

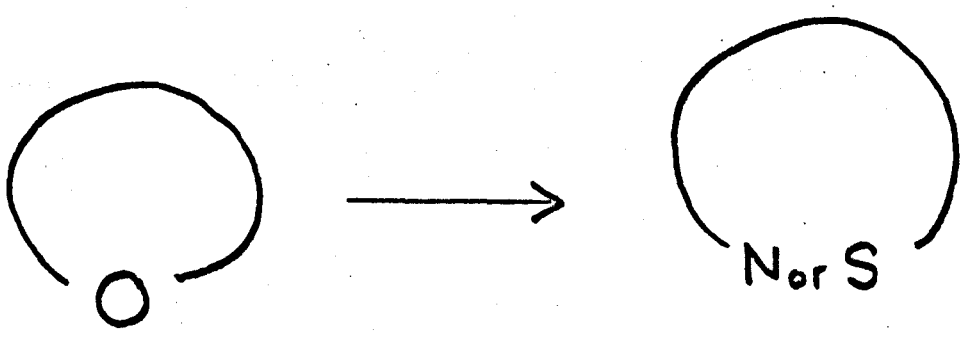
```
3
MIN 2 1 1 0 0 0
C 1 5
2 0 0 0? 1T
1 3 4 0U 3T
2 0 0 00 1F
2 5 0 0Q 2T
4 0 0 0? 1T
C 3 5
2 0 0 0? 1T
1 3 4 0U 3T
2 0 0 00 1F
2 5 0 0Q 2T
4 0 0 0? 1T
;
MIN 2 1 1 0 0 0
R 2
L6 AVUTJ            W
;
MIN 2 1 1 0 0 0
R 2
@L6 AVUTJ           W'
;
LOGIC 2
2 1 30
1 1 2A
```

Query deck for the reaction shown at the top of the Figure.



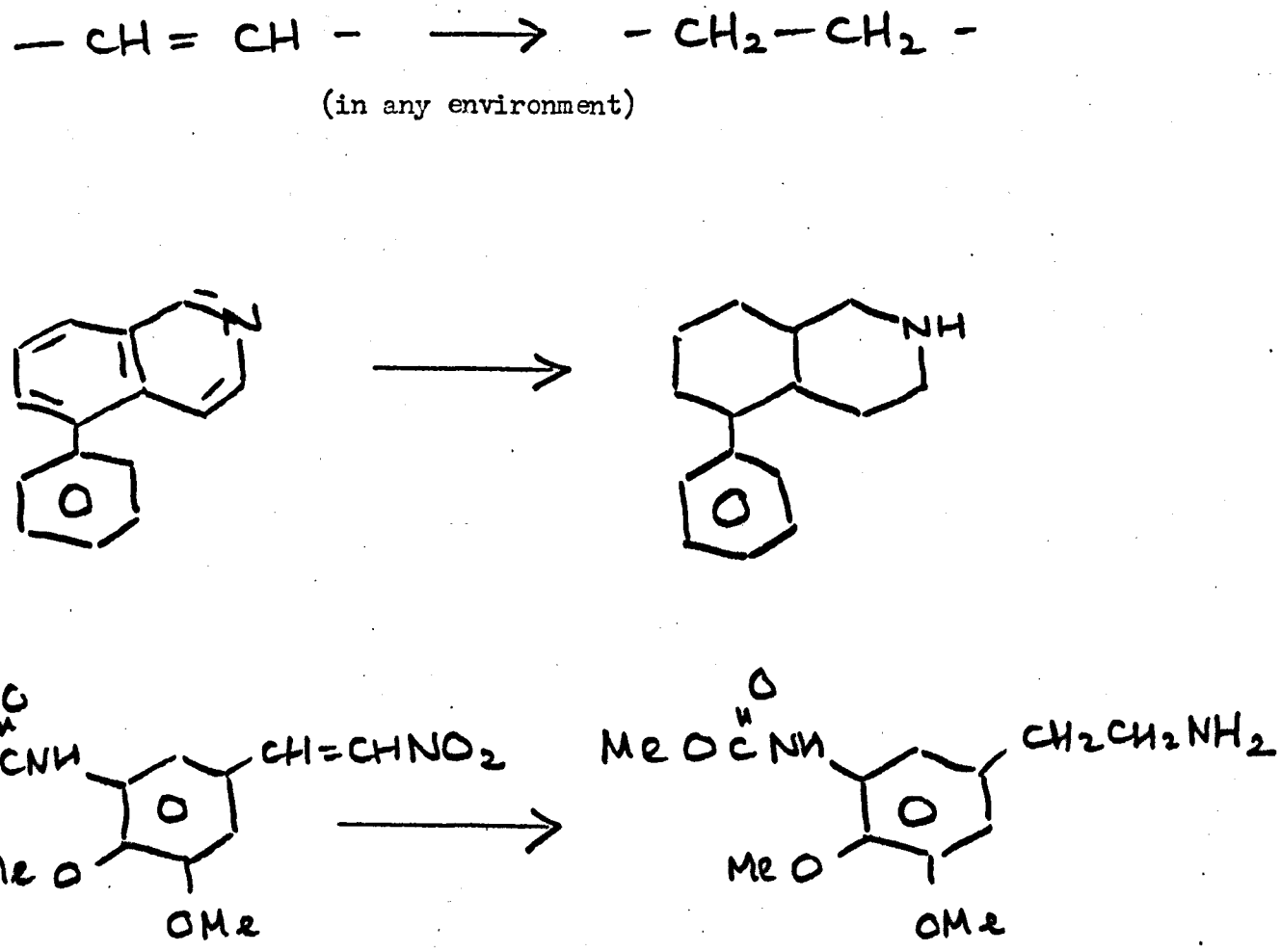

```
1
MIN 2 2 1 1 0 0
R 2
@T5 AMJ              W
R 2
@L5TJ                /
R 4
T8 AV BM EVUTJ       W
;
LOGIC 0
```

Example of a reaction in which ring statements are sufficient to define the query.




```
MIN 1 1 1 1 3 3
U 2 1QT
;
MIN 1 1 1 1 3 3
M 2 0 0 0 0 0 0 1 0 0
M 4 0 0 0 0 0 0 0 0 1
;
MIN 1 1 1 1 3 3
M 2 0 0 0 0 0 0 1 0 0
M 4 0 0 0 0 0 1 0 0 0
;
```

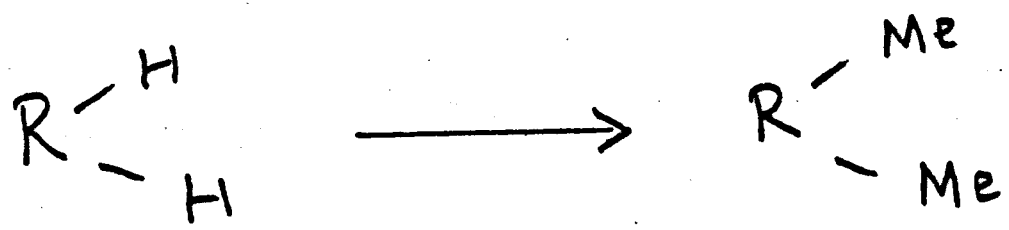
Use of U and M cards in query formulation

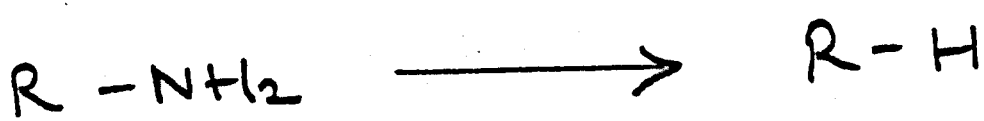


Examples of reactions judged as false drops to the query shown at the top of the Figure.

1	1	0	100.0	0.0
2	1	0	100.0	0.0
3	1	1	100.0	100.0
4	2	0	100.0	0.0
5	2	1	100.0	50.0
6	2	1	100.0	50.0
7	2	2	100.0	100.0
8	3	2	99.9	66.7
9	3	3	99.9	100.0
10	4	4	99.9	100.0
11	5	1	99.9	20.0
12	5	5	99.9	100.0
13	6	3	99.9	50.0
14	6	4	99.9	66.7
15	7	0	99.8	0.0
16	7	1	99.8	14.3
17	9	0	99.8	0.0
18	12	11	99.7	91.7
19	20	0	99.5	0.0
20	44	24	99.0	54.5
21	44	28	99.0	63.6
22	52	34	98.8	65.4
23	58	40	98.7	69.0
24	69	5	98.4	7.2
25	73	61	98.3	83.6
26	92	56	97.9	60.9
27	113	50	97.4	44.3

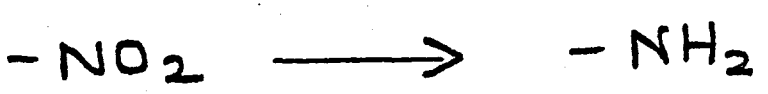
Retrieval results for the 27 queries that retrieved some material. The first figure in each set is the query number and this is followed by the number of reactions retrieved, the number of hits, the screenout and the precision.



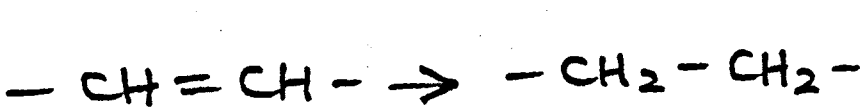


Two reaction queries which produced a large number of false drops.

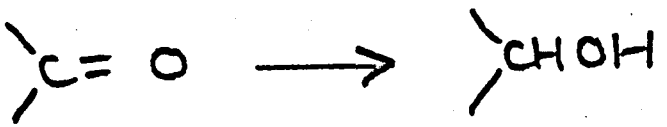
26



25



27



22

Examples of queries which produced large numbers of reactions. The figures beneath each reaction are the query numbers in Fig. V.13.

VI Conclusions and suggestions for future work

In this thesis we have described two methods for the automatic indexing of chemical reactions. Although designed for different types of structure representation, WLN and connection tables, and with different modes of use in mind, manual searching of printed indexes and mechanised searching of a serial bitstring file, they are based on a common principle. This is the identification of substructures in the reacting molecules which are as large as possible given the sole constraint that they must correspond to features present on both sides of the equation. Once these areas have been noted as common, the atoms or WLN symbols contained therein may be flagged in some way and the process repeated using the unmarked parts of the molecules until no further common areas may be found; the remaining portions of the molecules will then correspond to the reaction sites. In the connection table approach the common features are circular substructures which have been judged to be isomorphic using an approximate graph matching procedure based on an adaption of the Morgan algorithm; in the second method the identification of identical WLN symbol strings is used to determine the common features after the application of a multilevel fragmentation procedure.

The use of WLN implies a characterisation of reaction types on the basis of symbol, rather than substructural, differences but in many cases there is found to be a close correspondence between the two. This is due to the especial prominence given by the notation to those features which are of prime importance in synthetic work and thus an analysis based on WLN may be expected to give a simple and precise result for many reaction types. In other cases, however, there may be little or no similarity

between the reactant and product notations even though large parts of the molecules have not been changed in the course of the reaction. Also, the fact that a few symbols may represent quite large numbers of atoms and bonds implies both that a change may be described in somewhat generalised terms and that quite complicated symbol manipulations may be required in the course of the processing. Despite these limitations, the ability to provide character representations of the substructures involved allows one to produce printed indexes of reactions at very little cost which could be used in a manner similar to permuted WLN compound lists.

The analyses resulting from the connection table approach are only searchable in a wholly computerised system, access to the file being *via* a range of bit screens which allow the specification of a wide range of query requirements for both the reaction sites and the parent molecules. The need for computerised search is compensated for by the ability to carry out simultaneous substructure searches for both reacting and non-reacting features, by the variety of access modes provided, by the high degree of reaction site localisation and by the simplicity of the processing. The first of these features, dual access to both reacting and non-reacting substructures, is not available from the WLN analysis where the initial means of access is *via* the analysis fragments.

It is found that the two types of analysis are complementary in their coverage of the reactions in our file. Both methods deal satisfactorily with a wide range of acyclic transformations but ring changes are processed quite differently. The WLN analysis has been designed to isolate complete monocycles involved in a reaction whereas the connection table analysis

identifies just those individual ring atoms involved; the former approach is ideal for ring formation and cleavage reactions but insensitive to small changes within an individual ring whereas the converse applies to the latter approach. As ring changes account for at least 20% of the file studied, it can be seen that both analyses are needed if a comprehensive retrieval service is to be provided. The ring change information could be obtained using some form of ring perception algorithm but such techniques may prove quite expensive in terms of computer time whereas the WLNs of the ring systems in the reacting molecules, if available, may be processed very rapidly due to the fact that the smallest set of smallest rings has been previously isolated in the coding of the notation. The presence of the WLN symbol strings also provides a second level of search for those reactions which match the initial query bit string; such multilevel searching is common in industrial substructure search systems. Accordingly, the experimental reactions retrieval system that has been developed uses both types of analysis to characterise the reactions in the search file. The initial bit string descriptors are generated from the connection tables both of the reacting molecules and of the observed reaction sites, this permitting searches for highly specific changes.

The fragment screens used for the bit strings are obtained by the application of well established methods for textual character string manipulation to linear chemical substructure representations; this has resulted in a method of screen set generation and assignment which is computationally inexpensive, requires minimal manual intervention even at query formulation time and would seem to exhibit acceptable levels of retrieval

performance.

It is usual to include suggestions for further work arising out of that undertaken in the course of the thesis. However it is felt that there is little more that can be done in the present environment: the basic analysis algorithms have been shown to be practicable on the data available and such searches as have been carried out have produced acceptable results. Further development and evaluation would only seem worthwhile in the context of some specific external implementation. The screen set generation procedures, however, appear to be a useful tool in the development of general methodologies for screening systems and work in this area is continuing.

BIBLIOGRAPHY

Note: since very many of the references are to Journal of Chemical Documentation(latterly Journal of Chemical Information and Computer Science) the title of this journal has been abbreviated to J. in the list of references.

(1) Lynch, M. F., Computer-based information services in science and technology-principles and techniques, Peter Peregrinus Ltd., Stevenage(1974)

(2) Farmer, N. A., Schermer, C. A. and Wigington, R. L., The American Chemical Society composition system, CAS Report 5, 3-9(1976)

(3) Dammers, H. F., The economics of computer-based information systems; a review, J. Doc., 31(1), 38-45(1975)

(4) Leggate, P., Computer-based current awareness services, J. Doc., 31(2),93-115(1975)

(5) Barraclough, E. D., On-line searching in information retrieval, J. Doc. 33(3), 173-250(1977)

(6) Blick, A. R. and Magrill,D. S., The effect of the introduction of on-line facilities on the choice of search tools, Inf. Sci., 12(1), 32-37(1978)

(7) Sparck Jones, K., Automatic indexing, J. Doc., 30(4), 393-433(1974)

(8) Ash, J. E. and Hyde, E., Chemical information systems, Ellis Horwood, Chichester(1975)

(9) Wipke, W. T., Heller, S. R., Feldman, R. J. and Hyde, E., eds., Computer representation and manipulation of chemical information, John Wiley, New York(1973)

(10) Lynch, M. F., Harrison, J. M., Town, W. G. and Ash, J. E., Computer handling of chemical structure information, Macdonald, London(1971)

(11) Davis, C. H. and Rush, J. E., Information retrieval and documentation in chemistry, Greenwood Press, London(1974)

(12) Rouvray, D. H., The changing role of the symbol in the evolution of chemical notation, Endeav., 23-31(1977)

(13) Rush, J. E., Status of notation and topological systems and potential future trends, J., 16(4), 202-210(1976)

(14) Wiswesser, W. J., Historical development of chemical

notations, Ref. 8, 92-96

(15) National Academy of Sciences, Chemical structure information handling. A review of the literature 1962-1968, Nat. Acad. Sci., Washington(1969)

(16) Ash, J. E., Connection tables and their role in a system, Ref. 8, 156-176

(17) Smith, E. G. and Baker, P. A., The Wiswesser line-formula chemical notation, 3rd. ed.,Chemical Information Management Inc., New Jersey(1976)

(18) Baker, P. A., Palmer, G. and Nichols, P. W. L., The Wiswesser line-formula notation, Ref. 8,97-129

(19) Dittmar, P. G., Stobaugh, R. E. and Watson, C. E., The Chemical Abstracts Registry System. I. General design, J., 16(2), 111-121(1976)

(20) Maugh, T. H., Chemicals: how many are there?, Science, 199, 162, 1978

(21) Garfield, E., Revesz, G. S., Granito, C. E., Dorr, H. A., Caldern, M. M. and Warner, A., Index Chemicus Registry System: pragmatic approach to substructure chemical retrieval, J., 10(1), 54-58(1970)

(22) Stockton, F. G. and Merritt, R. L., The Shell chemical structure file system, J., 14(4), 166-170(1974)

(23) Tomea A. V. and Sorter, P. F., On-line substructure searching utilizing Wiswesser Line Notations, J., 16(4), 223-227(1976)

(24) Dammers, H. F. and Polton, D. J., Use of the IUPAC notation in computer processing of information on chemical structures, J., 8(3), 150-160(1968)

(25) Brown, H. D., Costlow, M., Cutler, F. A., Demott, A. N., Gall, W. B., Jacobus, D. P. and Miller, C. J., The computer-based chemical structure information system of Merck Sharp and Dohme

research laboratories, J., 16(1), 5-10(1976)

(26) Valls, J. and Schier, O., Chemical reaction indexing, Ref. 8, 243-258

(27) Valls, J., Reaction documentation, Ref. 9, 83-104

(28) Bawden, D., An investigation into the fundemental aspects of the analysis of chemical reactions for storage and retrieval, unpublished M. Sc. thesis, University of Sheffield(1974)

(29) Weyl, T. H.-, Die methoden der organischen Chemie, 3 vols., Thieme, Leipzig(1901-1911)

(30) Kresze, G., Present-day communication in chemistry-problems and possibilities, Angew. Chem. Int., 9(8), 545-550(1970)

(31) Meyer, E., Information science in relation to the chemists' needs, Ref. 8, 269-296

(32) Garfield, E., Revesz, G. S. and Batzig, J. H., The synthetic chemical literature from 1960 to 1969, Nature, 242, 307-309(1973)

(33) Price, D. J. de S., Little science, big science, Columbia University Press(1963)

(34) Hendrickson, J. B., Systematic synthesis design. IV. Numerical codification of construction reactions, J. Am. Chem. Soc., 97(20), 5784-5800(1975)

(35) Hendrickson, J. B., Systematic synthesis design, Topics in current chemistry, 62, 49-172(1976)

(36) Mockus, J., On the storage and retrieval of chemical reaction data, CAS internal report(1971)

(37) Lambourne, D. R., Reaction indexing, ICI Ltd.(Pharmaceuticals Division) internal report(1972)

(38) Schier, O., Nübling, W., Steidle, W. and Valls, J., A system for the documentation of chemical reactions, Angew. Chem. Int., 9(8), 599-604(1970)

(39) Osinga, M. and Stuart, A. A. V., Documentation of chemical reactions. I. A faceted classification, J., 13(1), 36-39(1973)

(40) Lynch, M. F., Nunn, P. R. and Radcliffe, J., Production of printed indexes of chemical reactions using Wiswesser line notations, J.(in press)

(41) Bersohn, M. and Esack, A., Computers and organic synthesis, Chem. Rev., 76(2), 269-282(1976)

(42) Orf, H. W., Computer-assisted synthetic analysis, Ph. D. thesis, Harvard University(1976)

(43) Bersohn, M. and Esack, A., A computer representation of synthetic reactions, Computers and Chemistry, 1(2), 103-107(1976)

(44) Sanders, A. F., Some applications of graph theory to the design of a heuristic program for the discovery of organic syntheses, Ph. D. thesis, State University of New York(1976)

(45) Gowan, J. E. and Wheeler, T. S., Name index of organic reactions, Longmans, London(1960)

(46) Denny, R. C., Named organic reactions, Butterworths, London (1969)

(47) Krauch, H. and Kunz, W., Organic name reactions, Wiley, London(1964)

(48) Vatsuro, K. V., Kalyatina, E. I., Karelin K. I. et al, Tezaurus organicheskikh reaktsii, Viniti, Moscow(1975)

(49) Clews, L. A., Characterisation of a reaction data-base as an aid to the evaluation of a chemical reaction retrieval system, unpublished M. Sc. thesis, University of Sheffield(1973)

(50) Patterson, A. M. and Bunnet, J. F., Systematic names for substitution reactions, Chem. Eng. News, 32, 4019(1954)

(51) Kunz, W., Informations-verarbeitung in einer Erfahrungswissenschaft, Nach. Dok., 314(18), 85-90(1967)

(52) Vleduts, G. E., Concerning one system of classification and

codification of organic reactions, Inform. Stor. Retr., 1(2/3), 117-146(1963)

(53) Ursprung-Fischer, R., New classification for organic reactions, Chem. Zeit., 89(24), 849-850(1965)

(54) Dyson, G. M. and Riley, E. F., Use of machine methods at Chemical Abstracts Service, J., 2(1), 19-22(1962)

(55) Organic syntheses, 56 vols., Wiley, New York(1921-1977)

(56) Mishchenko, G. L., Information retrieval in the field of reactions of organic chemistry, Zhurnal Vsesoyuznogo Khimicheskogo Obschestva im. D. I. Mendeleyeva, 16(1), 55-63(1971)

(57) Pothier, P. E., Substructure searching in CHEMLINE, Online, 1(2), 23-25(1977)

(58) Dunn, R. G., Fisanick, W. and Zamora, A., A chemical substructure search system based on Chemical Abstracts index nomenclature, J., 17(4), 212-219(1976)

(59) Rowland, J. F. B. and Veal, M. A., Structure-text and nomenclature-text searching for chemical information: an experiment with the Chemical Abstracts Integrated Subject File and Registry System, J., 17(2), 81-89(1977)

(60) Vander Stouw, G. G., Elliott, P. M. and Isenberg, A. C., Automated conversion of chemical substance names to atom-based connection tables, J., 14(4), 185-193(1974)

(61) Carpenter, N., Syntax-directed translation of organic chemical formulas into their two-dimensional representations, Computers and Chemistry, 1(1), 25-28(1976)

(62) Letter sent to J. F. Bunnet by the author(1977)

(63) Harrison, I. T. and Harrison, S., Compendium of organic synthetic methods, 2 vols., Wiley, New York(1971-1974)

(64) Buehler, C. A. and Pearson, D. E., Survey of organic syntheses, 2nd. ed., Wiley, New York(1977)

(65) Jones, R., Polko, R., Seitz, G. and Unger, R., Synthetica Merck: Reagenzien für die organische Synthese, Merck, Darmstadt (1969)

(66) Fieser, M. and Fieser, L. F., Reagents for organic synthesis, 5 vols., Wiley, New York(1967-1975)

(67) Weygand, C., Organisch-chemische Experimentierkunst, 3 vols., Barth, Leipzig(1938)

(68) Theilheimer, W., Synthetic methods of organic chemistry, 27 vols., Karger, Basel(1946-1975)

(69) Evans, P. N., Organization of information on organic reactions, unpublished M. Sc. thesis, University of Sheffield(1967)

(70) Mathieu, J. and Allais, A., Cahiers de synthese organique, 12 vols., Masson, Paris(1957-1966)

(71) Derwent Publications, Chemical Reactions Documentation Service

(72) Bawden, D. and Jackson, F., personal communication(1977)

(73) Ziegler, H., A new information system for organic reactions, J., 6(2), 81-89(1966)

(74) Vleduts, G. E., Mishchenko, G. L., Rokhlin, E. M. and Tarasova, L. V., Experimental development and improvement of a factographic information-retrieval system for organic chemistry(for example, an information retrieval system for the chemistry of fluoroorganic compounds), in Chem. Abs., 70, 25535m(1969)

(75) Eakin, D. R. and Hyde, E., Evaluation of on-line techniques in a sub-structure search system, Ref. 9,1-30

(76) Eakin, D. R. and Warr, W. A., Computerised aids to organic synthesis in a pharmaceutical research company, Am. Chem. Soc. Symposium Series(in the press)

(77) Seddon, J. M., An evaluation of chemical reaction analysis

and retrieval systems, unpublished M. Sc. thesis, University of Sheffield(1973)

(78) Osinga, M. and Stuart, A. A. V., Documentation of chemical reactions. II. Analysis of the Wiswesser line notation, J., 14(4), 194-198(1974)

(79) Meyer, E., The IDC system,for chemical documentation, J., 9(2), 109-113(1969)

(80) Lobeck, M. A., The use of the IDC system, Angew. Chem. Int., 9(8), 576-583(1970)

(81) Rossler, S. and Kolb, A., The GREMAS system, an integral part of the IDC system for chemical documentation, J., 10(2), 128-134 (1970)

(82) Gelberg, A., Rapid structure searches via permuted chemical line notations. IV. A reactant index, J., 6(1), 60-61(1966)

(83) Shevyakova, L. A. and Stoyanovich, F. M., Information retrieval system for the chemistry of organosulphur compounds, Nauk. Tekh. Infor., 12(1), 21-22(1965)

(84) Meyer, E., Mechanization of chemical documentation, Angew. Chem. Int., 4(4), 347-352(1965)

(85) Urbankova, I., System for indexing organic chemical reactions, Chem. Zvesti, 30(5), 715-718(1976)

(86) Hendrickson, J. B., Cram, D. J. and Hammond, G. S., Organic chemistry, 3rd. ed., McGraw-Hill, New York(1970)

(87) Guthrie, R. D., A suggestion for the revision of mechanistic designations, J. Org. Chem., 40(4), 402-407(1975)

(88) Mathieu, J., Allis, A. and Valls, J., Nucleofuger und elektrofuger Austritt, Angew. Chem., 72(2), 71-74(1960)

(89) Satchell, D. P. N., The classification of chemical reactions, Naturwissenschaften, 64(3), 113-121(1977)

(90) International Union of Pure and Applied Chemistry, Minutes

of the fifth meeting of the commision on physical organic chemistry, Warsaw(1977)

(91) Mishchenko, G. P., Vleduts, G. E. and Shefter, A. M., Automatic indexing of reactions in an information-retrieval systems for organic chemistry, Nauk. Tekh. Inform., 10(1), 13-17(1965)

(92) Crowe, J. E., Lynch, M. F. and Town, W. G., Analysis of structural characteristics of chemical compounds in a large computer-based file. Part I: non-cyclic fragments, J. Chem. Soc.(C), 990-996(1970)

(93) Mishchenko, G. L., Empirical formulas of bonds of compounds and their possible role in retrieving factographic information in chemistry, Inf. Probl. Sovrem. Khim., 25-38(1976)

(94) Armitage, J. E. and Lynch, M. F., Automatic detection of structural similarities among chemical compounds, J. Chem. Soc.(C), 521-528(1967)

(95) Armitage, J. E., Crowe, J. E., Evans, P. N., Lynch, M. F. and McGuirk, J. A., Documentation of chemical reactions by computer analysis of structural changes, J., 7(4), 209-215(1967)

(96) McGuirk, J. A., Computer synthesis of acyclic fragments of organic compounds, unpublished M. Sc. thesis, University of Sheffield(1967)

(97) Cone, M. M., Venkataraghavan, R. and McLafferty, F. W., Molecular structure comparison program for the identification of maximal common substructures, J. Am. Chem. Soc., 99(23), 7668-7671(1977)

(98) Harrison, J. M. and Lynch, M. F., Computer analysis of chemical reactions for storage and retrieval, J. Chem. Soc.(C), 2082-2087(1970)

(99) Lynch, M. F., Computer organisation and retrieval of chemical structure information, Office of Scientific and Technical Information

report no. 5203

(100) Campbell, M. B. M., Evaluation of a chemical reactions search and retrieval system, unpublished M. Sc. thesis, University of Sheffield(1972)

(101) Clinging, R. and Lynch, M. F., Production of printed indexes of chemical reactions. I. Analysis of functional group interconversions, J., 13(2), 98-102(1973)

(102) Clinging, R. and Lynch, M. F., Production of printed indexes of chemical reactions. II. Analysis of reactions involving ring formation, cleavage and interconversion, J., 14(2), 69-71(1974)

(103) Crowe, J. E., Leggate, P., Rossiter, B. N. and Rowland, J. F. B., The searching of Wiswesser line notations by means of a character-matching serial search, J., 13(2), 85-92(1973)

(104) Lynch, M. F., Nunn, P. R. and Radcliffe, J., Development and assessment of an automatic system for analysing chemical reactions, British Library Research and Development Department report no. 5236

(105) Nunn, P. R., The automatic analysis of chemical reactions, unpublished M. Sc. thesis, University of Sheffield(1974)

(106) Vickery, B. C., Faceted classification: a guide to the construction and use of special schemes, ASLIB, London(1960)

(107) Osinga, M. and Stuart, A. A. V., Documentation of chemical reactions. III. Encoding the facets, J., 16(3), 165-171(1976)

(108) Adamson, G. W., Lynch, M. F. and Town, W. G., Analysis of structural characteristics of chemical compounds in a computer-based file. Part II. Atom-centred fragments, J. Chem. Soc.(C), 3702-3706(1971)

(109) Hyde, E., Matthews, F. W., Thomson, L. H. and Wiswesser, W. J., Conversion of Wiswesser notation to a connectivity matrix for organic compounds, J., 7(4), 200-204(1967)

(110) Thomson, L. H., Hyde, E. and Matthews, F. W., Organic search and display using a connectivity matrix derived from the Wiswesser notation, J., 7(4), 204-207(1967)

(111) Granito, C. E., CHEMTRAN and the interconversion of chemical substructure systems, J., 13(2), 72-74(1973)

(112) Granito, C. E., Roberts, S. and Gibson, G. W., The conversion of Wiswesser line notations to Ring codes. I. The conversion of ring systems, J., 12(3), 190-196(1972)

(113) Granito, C. E., The conversion of WLN to Ring Code, Paper given at the ICRS User's meeting, London(1976)

(114) Vleduts, G. E., Development of a combined WLN/CTR multilevel approach to the algorithmical analysis of chemical reactions in view of their automatic indexing, British Library Research and Development report no. 5399, London(1977)

(115) Unger, S. H., GIT-a heuristic program for testing pairs of directed line graphs for isomorphism, Comm. Ass. Comp. Mach., 7(1), 26-34(1964)

(116) Sussenguth, E. H., A graph-theoretic algorithm for matching chemical structures, J., 5(1), 36-43(1965)

(117) Corneil, D. G. and Gotlieb, C. C., An efficient algorithm for graph isomorphism, J. Ass. Comp. Mach., 17(1), 51-64(1970)

(118) Berztiss, A. T., A backtrack procedure for isomorphism of directed graphs, J. Ass. Comp. Mach., 20(3), 365-377(1973)

(119) Ullman, J. R., An algorithm for subgraph isomorphism, J. Ass. Comp. Mach., 23(1), 31-42(1976)

(120) Levi, G., A note on the derivation of maximal common subgraphs of two directed or undirected graphs, Calcolo, 9(1), 1-12(1972)

(121) Tarjan, R. E., Graph algorithms in chemical computation,

in Algorithms for chemical computations, Am. Chem. Soc. Symposium Series 46, 1-19(1977)

(122) Barrow, H. G. and Burstall, R. M., Subgraph isomorphism, matching relational structures and maximal cliques, Inf. Proc. Lett., 4(4), 83-84(1976)

(123) Ming, T. K. and Tauber, S. J., Chemical structure and substructure search by set reduction, J., 11(1), 47-51(1971)

(124) Figueras, J., Substructure search by set reduction, J., 12(4), 237-244(1972)

(125) Lynch, M. F., Screening large chemical files, Ref. 8, 177-194

(126) Bawden, D., Substructural analysis techniques for structure-property within computerised chemical information systems, unpublished Ph. D. thesis, University of Sheffield(1977)

(127) Wipke, W. T. and Dyott, T. M., Use of ring assemblies in a ring perception algorithm, J., 15(3), 140-147(1975)

(128) Zamora, A., An algorithm for finding the smallest set of smallest rings, J., 16(1), 40-43(1975)

(129) Eakins, J. P., An analysis of the distribution of six-membered ring fragments in a random file of chemical structures, unpublished M. Sc. thesis, University of Sheffield(1970)

(130) Adamson, G. W., Creasey, S. E., Eakins, J. P. and Lynch, M. F., Analysis of structural characteristics of chemical compounds in a large computer-based file. Part V. More detailed cyclic fragments, J. Chem. Soc.(C), 2071-2076(1973)

(131) Bedrosian, S. D. and Milne, M. B., Graphical representation for automated retrieval of a class of fused six-rings, J., 17(1), 47-49(1977)

(132) Palmer, G., WLN ring decoder library routine, unpublished report, ICI Ltd.(Pharmaceuticals Division)(1974)

(133) Leo, A., Elkins, D. and Hansch, C., Computerized management of structure-activity data. Part III. Computerized decoding and manipulation of ring structures coded in WLN, J., 14(2), 65-69(1974)

(134) Walker, J. C. and Tauber, S. J., Connection tables from Wiswesser Line notation: a partial algorithm, Proceedings of the Army Chemical Information and Data Systems program ed. Mitchell, J. P., AD 665 397(1968)

(135) Lefkowitz, D., A chemical notation and code for computer manipulation, J., 7(4), 186-192(1967)

(136) Milne, M., Lefkowitz, D., Hill, H. and Powers, R., Search of CA Registry(1.25 million compounds) with the Topological Screen System, J., 12(3), 183-189(1972)

(137) King, D. R., The generation of eight-atom fragments for use as screens in the sub-structure searching of files of chemical compounds, unpublished M. Sc. thesis, University of Sheffield(1971)

(138) Gannon, M. T., unpublished results

(139) Garagnarni, E. and Bart, J. C. J., Organic reaction schemes and general reaction-matrix types. Part III. A quantitative analysis, Z. Naturforsch., 32B, 465-468(1977)

(140) Lynch, M. F., The micrstructure of chemical data-bases and the choice of representation for retrieval, Ref. 9, 31-54

(141) Morgan, H. L., The generation of a unique machine-description for chemical structures - a technique developed at Chemical Abstracts Service, J., 5(2), 107-113(1965)

(142) Lynch, M. F., personal communication

(143) Wipke, W. T. and Dyott, T. M., Stereochemically unique naming algorithm, J. Am. Chem. Soc., 96(15), 4834-4842(1974)

(144) Morris, R., Scatter storage techniques, Comm. Ass. Comp. Mach., 11(1), 38-44(1968)

(145) Maurer, W. D. and Lewis, T. G., Hash table methods, Computing Surveys, 7(1), 5-19(1975)

(146) Feldman, R. J. and Heller, S. R., An application of interactive graphics-the nested retrieval of chemical structures, J., 12(1), 48-54(1972)

(147) Feldman, R. J., Milne, G. W. A., Heller, S. R., Fein, A., Miller, J. A. and Koch, B., An interactive substructure serach system, J., 17(3), 157-163(1977)

(148) Feldman, R. J., paper given at the C. N. A.(UK) meeting, Warrington(1978)

(149) Machin, P. A., Froud, D. and Elder, M., CSSR Crystal Structure Search Retrieval, Science Research Council, Daresbury(1977)

(150) Evans, L. A., Lynch, M. F. and Willett , P., Structural search codes for on-line compound registration, J., (in press)

(151) O'Korn, L. J., personal communication(1977)

(152) Freeland, R. G., Funk, S. J., O'Korn, L. J. and Wilson, G. A., Augmented connectivity molform - a technique for recognition of structure topology, Abstract CHLT 29, 169th Am. Chem. Soc. meeting, Philadelphia(1975)

(153) Dubois, J. E., Ordered chromatic graphs and limited environment concepts, in Balaban, A. T., ed., Chemical applications of graph theory, Academic Press, London(1976),

(154) Wiswesser, W. J., The "dot-plot" computer program, Proceedings of the Army Chemical Information and Data Systems program ed. Mitchell, J. P., AD 665 397(1968)

(155) Shelley, C. A. and Munk, M. E., Computer perception of topological symmetry, J., 17(2), 110-113(1977)

(156) Schmidt, D. C. and Druffel, L. E., A fast backtracking algorithm to test directed graphs for isomorphism using distance matrices, J. Assoc. Comp. Mach., 23(3), 433-445(1976)

(157) Cossum, W. E., Krakiwsky, M. L. and Lynch, M. F., Advances

in automatic chemical substructure searching techniques, J., 5(1), 33-35(1965)

(158) Ray, L. C. and Kirsch, R. A., Finding chemical records by digital computers, Science, 126, 814-819(1957)

(159) Wipke, W. T., Krishna, S. and Ouchi, G. I., Hash functions for rapid storage and retrieval of chemical structures, J., 18(1), 32-36(1978)

(160) Woodward, P. M. and Bond, S. G., Algol68-R users guide, HMSO, London(1974)

(161) Golomb, S. W. and Baumert, L. D., Backtrack programming, J. Assoc. Comp. Mach., 12(4), 516-524(1965)

(162) McGregor, J. J., personal communication(1978)

(163) Evans, L. A., Topological indexes and their application to chemical structure registration, unpublished M. Sc. thesis, University of Sheffield(1977)

(164) Craig, P. N. and Ebert, H. M., Eleven years of structure retrieval using the SK&F fragment codes, J., 9(3), 141-146(1969)

(165) Sorter, P. F., Granito, C. E., Gilver, J. C., Gelberg, A., Williams, R. J. and Metcalf, E. A., Rapid structure searches via permuted chemical line-notations, J., 4(1), 56-60(1964)

(166) Granito, C. E., Becker, G. T., Roberts, S., Wiswesser, W. J. and Windlinx, K. J., Computer-generated substructure codes (bit screens), J., 11(2), 106-110(1971)

(167) Lewis, H. R. and Papadimitriou, C. H., The efficiency of algorithms, Scientific American, 238(1), 96-109(1978)

(168) Feldman, A. and Hodes, L., An efficient design for chemical structure searching. I. The screens, J., 15(3), 147-152(1975)

(169) Smith, D. H., Masinter, L. M. and Sridharan, Heuristic DENDRAL: analysis of molecular structure, Ref. 9, 287-315

(170) Mooers, C. N., Zatocoding applied to the mechanical organization of knowledge, Am. Doc., 2(1), 20-32(1951)

(171) Brookes, B. C., The complete Bradford-Zipf bibliography,

J. Doc., 25(1), 58-60(1969)

(172) Fairthorne, R. A., Empirical hyperbolic distributions (Bradford-Zipf-Mandelbrot) for bibliometric description and prediction, J. Doc., 25(4), 319-343(1969)

(173) Salton, G., Dynamic information and library processing, Chapter 5, Prentice-Hall, New Jersey(1975)

(174) Bird, P. R., Some sampling characteristics of bibliometric distributions, J. Informatics, 1(2), 69-80(1977)

(175) Adamson, G. W., Cowell, J., Lynch, M. F., Town, W. G. and Yapp, A. M., Analysis of structural characteristics of chemical compounds in a large computer-based file. Part IV. Cyclic fragments, J. Chem. Soc.(C), 863-865(1973)

(176) Adamson, G. W., Clinch, V. A. and Lynch, M. F., Relationship between query and data-base microstructure in general substructure search systems, J., 13(3), 133-136(1973)

(177) Shannon, C. E., A mathematical theory of communication, Bell System Technical Journal, 27(3), 379-423 and 27(4), 623-656 (1948)

(178) Adamson, G. W., Lambourne, D. R. and Lynch, M. F., Analysis of structural characteristics of chemical compounds in a large computer-based file, Part III. Statistical association of fragment incidence, J. Chem. Soc.(C), 2428-2433(1972)

(179) Adamson, G. W., Cowell,J,Lynch, M. F., McLure, A. H. W., Town, W. G. and Yapp, A. M., Strategic considerations in the design of a screening system for substructure searches of chemical structure files, J., 13(3), 153-157(1973)

(180) Hodes, L., Selection of descriptors according to discrimination and redundancy. Application to chemical structure searching, J., 16(2), 88-92(1976)

(181) Hodes, L. and Feldman, A., An efficient design for chemical structure searching. Part II. The file organisation, J., 18(2),

96-100(1978)

(182) Salton, G., A theory of indexing, Regional conference series in applied mathematics, S.I.A.M.(1975)

(183) Salton, G., Wong, A. and Yang, C. S., A vector space model for automatic indexing, Comm. Ass. Comp. Mach., 18(11), 613-620(1975)

(184) Bersohn, M., Rapid generation of reactants in organic synthesis programs, Am. Chem. Soc. Symposium Series(in the press)

(185) Jochum, C. and Gasteiger, J., Canonical numbering and constitutional symmetry, J., 17(2), 113-117(1977)

(186) Eakin, D. R. and Burgess, M. T., The CROSSBOW Mark II connection table, internal report, ICI Ltd.(Pharmaceuticals Division)(1975)

(187) Adamson, G. W., Clinch, V. A., Creasey, S. E. and Lynch, M. F., Distribution of fragment representations in a chemical substructure search screening system, J., 14(2), 72-74(1974)

(188) Barnard, J. M., unpublished results

(189) Schuegraf, E. J. and Heaps, H. S., Selection of equifrequent word fragments for information retrieval, Inf. Stor. Retr., 9(12), 697-711(1973)

(190) Williams, P. W., Criteria for choosing subsets to obtain maximum relative entropy, Computer J., 21(1), 57-62(1978)

(191) Brack, E. V., Cooper, D. and Lynch, M. F., The stability of symbol sets produced by variety generation from bibliographical data, Program, 12(2), 64-77(1978)

(192) Cooper, D. and Lynch, M. F., The compression of Wiswesser Line Notations using variety generation, in preparation

(193) Fugmann, R., The IDC system, Ref. 8, 195-226

(194) Eakin, D. R., The ICI CROSSBOW system, Ref. 8, 227-242

(195) Eakin, D. R. and Palmer, G., The CROSSBOW fragment screen

automatically generated from the Wiswesser Line Notation, internal report, ICI Ltd.(Pharmaceuticals Division)(1973)

(196) Clare, A. C., Cook, E. M. and Lynch, M. F., The identification of variable-length, equifrequent character strings in a natural language data base, Computer J., 15(3), 259-262(1972)

(197) Yeates, A. R., Text compression in the Brown corpus, unpublished M. A. thesis, University of Sheffield(1977)

(198) Powers, R. V. and Hill, H. N., Designing CIDS - the U.S. Army Chemical Information and Data System, J., 11(1), 30-38(1971)

(199) Brack, E. V., personal communication

(200) Polton, D. J., A computer process for substructure searches on compound structures ciphered in the IUPAC notation, Inf. Stor. Retr., 8(4), 191-201(1972)

(201) Powell, J., Rowland, J. F. B. and Veal, M. A., Correlated structure-text searching for chemical information, British Library Research and Development Department report no. 5292(1976)

(202) Granito, C. E. and Rosenberg, M. D., Chemical Substructure Index(CSI) - a new research tool, J., 11(4), 251-256(1971)

(203) Sheng, A., Lupi, L., Ronayne, M., Sprules, A. and Zornetzer, S., Hoffman-La Roche's on-line/batch interactive chemical information system, J., 14(4), 179-184(1974)

(204) Schultz, J. L., Handling chemical information in the Du Pont Central Report Index, J., 14(4), 171-179(1974)

(205) Schenk, H. R. and Wegmüller, F., Substructure search by means of the Chemical Abstracts Service Chemical Registry II System, J., 16(3), 153-161(1976)

(206) Zamora, A. and Dayton, D. L., The Chemical Abstracts Service Chemical Registry System. V. Structure input and editing, J., 16(4), 219-222(1976)

(207) International Computers Limited, Operating systems GEORGE 3 and 4, 5th ed., International Computers Limited, London(1976)

(208) Corey, E. J., Cramer, R. D. and Howe, W. J., Computer-assisted synthetic analysis for complex molecules. Methods and procedures for machine generation of synthetic intermediates, J. Am. Chem. Soc., 94(2), 440-459(1972)

(209) Van Rijsbergen, C. J., Information retrieval, Butterworth, London(1975)

(210) Salton, G., The SMART system - experiments in automatic document processing, Prentice-Hall, New Jersey(1971)

(211) Cleverdon, C. W., On the inverse relationship of recall and precision, J. Doc., 28(3), 195-201(1972)

(212) Adamson, G. W., Bush, J. A., McLure, A. H. W. and Lynch, M. F., An evaluation of a substructure search screen system based on bond-centred fragments, J., 14(1), 44-48(1974)

APPENDIX I

```
 0  SCREENSETGENERATION 'WITH' MTH 'FROM' ;LIB.SHEF68ALBUM
 1  'BEGIN'
 2
 3            'C' AUTHOR PETER WILLETT , PGSLIS ; 1977.
 4            THIS PROGRAM READS AN INPUT TAPE FILE CONTAINING POTENTIAL SCREENS
 5            RECORDS.  THE FILE HAS BEEN SORTED FIRSTLY UPON THE SIZE OF THE
 6            FRAGMENT INTEGER STRINGS AND SECONDLY UPON THEIR INCIDENCE.  THE
 7            SINGLE INTEGER STRINGS ARE READ INTO THE ARRAY "ALPHAFILE" WHICH IS
 8            TWICE THE SIZE OF THE DESIRED SCREEN SET; IT IS ASSUMED THAT THIS SIZE
 9            IS GREATER THAN THE NUMBER OF DIFFERENT SINGLE INTEGER STRINGS.
10            FOR SUBSEQUENT GROUPS OF STRINGS OF A GIVEN SIZE THE PROCEDURE IS :
11            (1) READ A STRING
12            (2) CHECK THAT THE STRING INCIDENCE IS GREATER THAN THE
13            THRESHOLD VALUE
14            (3) CHECK ALSO THAT THE PARENT STRING IS PRESENT IN THE CURRENT
15            SCREEN SET WHICH IS STORED IN "ALPHAFILE" IN ALPHABETICAL
16            ORDER
17            (4) STORE THE STRING IN THE NEXT LOCATION IN "KEYSET" WHICH
18            IS KEPT IN DECREASING INCIDENCE ORDER
19            (5) AT THE END OF THE STRINGS OF A GIVEN LENGTH, THE ARRAY "ALPHAFILE"
20            , CONTAINING THE CURRENT SCREEN SET IS MERGED WITH "KEYSET" AND
21            THEN SORTED INTO ALPHABETICAL ORDER.  THE POTENTIAL NEW SCREENS, IE
22            THOSE OF THE CURRENT SIZE, ARISING FROM A GIVEN PARENT STRING ARE
23            ADDED TO THE SET IF THERE IS A SUFFICIENT DIFFERENCE IN INCIDENCE
24            BETWEEN THEM AND THE PARENT
25            (6) THE TOP HALF OF "ALPHAFILE" BECOMES THE CURRENT SCREEN SET. 'C'
26
27       'MTFDA' TAPEFILE ; 'CLEAR' TAPEFILE ;
28       'PROC' EXCEP := 'VOID' :
29         (FAULT("TAPE ERROR")) ;
30       MTDEF(1,1,512,TAPEFILE,EXCEP,1) ;
31       'TO' 15 'DO' NEXTLINE(STANDOUT) ;
32       'CHARPUT' SCREENSOUT ; OPENC(SCREENSOUT,LINE PRINTER , 1) ;
33            'C' MAXLENGTH IS THE MAXIMUM NUMBER OF INTEGERS IN THE SCREEN ;
34            KS THE DESIRED SCREEN SET SIZE.  THRESHOLD IS CALCULATED FROM THE
```

```
          FIRST INPUT TAPE RECORD WHOSE FINAL WORD CONTAINS THE TOTAL NUMBER OF
          INTEGER STRINGS SUBMITTED TO PROGRAM "SUMMPROG" 'C'
'INT' KS ; READ(KS) ;
'INT' S := 0 ; LASTSETSIZE := 0 ;
[0:50]'BYTES' TAPERECORD ;
MTRD(1,TAPERECORD[0]) ;
'INT' MAXLENGTH := ('ABS' TAPERECORD[0])=3 ;
'INT' TOTALINCIDENCE = 'ABS' TAPERECORD[MAXLENGTH+2] ;
'INT' THRESHOLD = TOTALINCIDENCE'/'(4*(KS+1)) ;
'INT' KS2 = KS + KS ;
'MODE''KEY' =
  'STRUCT'('INT' SIZE , [1:MAXLENGTH]'INT' TYPE , 'INT' TOTAL,NONREDUNDANT);
[1:KS2]'KEY' ALPHAFILE ; [1:KS]'KEY' KEYSET ; 'KEY' CURRENT ;
'REF''INT' LIMITKS = TOTAL 'OF' ALPHAFILE[KS] ;
'REF''INT' TOTALKS = TOTAL 'OF' KEYSET[KS] ;
'REF'[]'BYTES' RECORD = TAPERECORD[0:MAXLENGTH+2] ;

'OP''EQUALS' = ('REF'[]'INT' A,B)'BOOL' :
  'BEGIN'
    'BOOL' SAME := 'TRUE' ;
    'FOR' X 'TO''UPB' A 'WHILE' SAME 'DO' SAME := A[X] = B[X] ;
    SAME
  'END' ;

'OP''LESSTHAN' = ('REF'[]'INT' A,B)'BOOL' :
  'BEGIN'
    'INT' X := 0 ; 'BOOL' SAME := 'TRUE' ;
    'TO''UPB' A 'WHILE' SAME 'DO' (X 'PLUS' 1 ; SAME := A[X] = B[X] ) ;
    'IF' A[X] < B[X] 'THEN''TRUE''ELSE''FALSE''FI'
  'END' ;

'OP''SBA' = ('INT' I)'BYTES' :
  ('BYTES' B := 'BYTES''CODE' 0,5/I 'EDOC' ; B) ;

'PROC' FINDPARENT = ('INT' M) :
```

```
    'BEGIN'
      'C' DOES A BINARY SEARCH OF THE ALPHAORDERED SCREEN SET OBTAINED
      AT THE END OF THE PREVIOUS ITERATION 'C'
    'INT' L := 0 , R := LASTSETSIZE+1 , J := 'ENTIER'((L+R)/2) ;
    'INT' FOUND := 0 ;
    'REF'[]'INT' A = (TYPE 'OF' CURRENT)[1:M] ;
    'WHILE' J # L 'DO'
      'BEGIN'
        'REF'[]'INT' TAEJ = TYPE 'OF' ALPHAFILE[J] ;
        'IF' A 'EQUALS' TAEJ[1:M] 'THEN' FOUND := TOTAL 'OF'
          ALPHAFILE[J] ; L := J
        'ELSE' A 'LESSTHAN' TAEJ[1:M] 'THEN' R := J ;
          J := 'ENTIER'((L+R)/2)
        'ELSE' L := J ; J := 'ENTIER'((L+R)/2) 'FI'
      'END' ;
    'IF' FOUND = TOTAL 'OF' CURRENT > THRESHOLD 'THEN'
      KEYSET[S 'PLUS' 1] := CURRENT
    'FI'
  'END' ;

'PROC'('REF'[]'KEY' , 'PROC'('REF''KEY' , 'REF''KEY')'BOOL') ORDER
  = SHELLORDER 'AS' ORDER ;
'PROC' AUXORDER = ('REF''KEY' U , V)'BOOL' :
  'BEGIN'
    'KEY' W ;
    'IF' TOTAL 'OF' U < TOTAL 'OF' V 'THEN' W := U ; U := V ; V := W ;
      'TRUE''ELSE''FALSE''FI'
  'END' ;
'PROC' ALPHAORDER = ('REF''KEY' U,V)'BOOL' :
  'BEGIN'
    'KEY' W ;
    'IF' TYPE 'OF' U 'LESSTHAN' TYPE 'OF' V 'THEN''FALSE'
      'ELSE' W := U ; U := V ; V := W ; 'TRUE''FI'
  'END' ;
'PROC' SIZEORDER = ('REF''KEY' U,V)'BOOL' :
```

```
105         'BEGIN'
106           'KEY' U ; 'IF' SIZE 'OF' U < SIZE 'OF' V 'THEN''FALSE''ELSE'
107           W := U ; U := V ; V := W ; 'TRUE''FI'
108         'END' ;
109
110       'PROC' MAXLENGTHPRESENT = 'INT' :
111         'BEGIN'
112           'INT' MAX := 0 ;
113           'FOR' X 'TO' LASTSETSIZE 'DO'
114             'IF' SIZE 'OF' ALPHAFILE[X] > MAX 'THEN'
115             MAX := SIZE 'OF' ALPHAFILE[X] 'FI' ;
116           MAX
117         'END' ;
118
119       'PROC' UPDATEALPHAFILE = ('INT' SCREENLENGTH)'VOID' :
120         'BEGIN'
121             'C' AT THE END OF THE SCREENS OF A GIVEN LENGTH , THE POSSIBLE NEW
122             SCREENS IN "KEYSET" ARE MERGED WITH THE SET OBTAINED AT THE END OF THE
123             LAST SCREEN-LENGTH , IN "ALPHAFILE" , AND THE NEW SET , LASTSETSIZE
124             LARGE, IS SORTED INTO DESCENDING FREQUENCY ORDER 'C'
125           'BOOL' B := 'TRUE' ;
126           'INT' LSS := LASTSETSIZE ;
127           'FOR' Y 'TO' S 'WHILE' LSS < KS2 'DO' ALPHAFILE[LSS 'PLUS' 1] :=
128             KEYSET[Y] ;
129           ORDER(ALPHAFILE[1:LSS],ALPHAORDER) ;
130           'INT' SCREENSDELETED := 0 ;
131           'FOR' X 'TO' LSS-1 'DO'
132             'IF' SIZE 'OF' ALPHAFILE[X] = SCREENLENGTH 'THEN'
133               'REF''KEY' AX = ALPHAFILE[X] ;
134               'INT' SAX = SIZE 'OF' AX + 1 ;
135               'INT' T := TOTAL 'OF' AX ;
136               'INT' Y := X+1 ;
137               'WHILE' Y <= LSS 'AND' SIZE 'OF' ALPHAFILE[Y] = SAX 'DO'
138                 (T 'MINUS' TOTAL 'OF' ALPHAFILE[Y] ; Y 'PLUS' 1) ;
139               Y 'MINUS' 1 ;
```

```
140             'C' OFFSPRING SCREENS ARE DELETED ONE BY ONE , IN INVERSE
141             FREQUENCY ORDER, UNTIL THE NONREDUNDANT PARENTAL FREQUENCY
142             RISES ABOVE THE THRESHOLD 'C'
143             'IF' TOTAL 'OF' AX <= THRESHOLD 'THEN''FOR' Z 'FROM' X 'TO' Y 'DO'
144               TOTAL 'OF' ALPHAFILE[Z] := 0
145             'ELSE'
146             'WHILE' T <= THRESHOLD 'DO''BEGIN'
147                 'INT' SMALLEST := 1 ; MIN := 999999 ;
148                 'FOR' Z 'FROM' X+1 'TO' Y 'DO'
149                   'IF' TOTAL 'OF' ALPHAFILE[Z] < MIN 'THEN'
150                     SMALLEST := Z ; MIN := TOTAL 'OF' ALPHAFILE[Z] 'FI' ;
151                 T 'PLUS' MIN ;
152                 TOTAL 'OF' ALPHAFILE[SMALLEST] := 0 ;
153                 SCREENSDELETED 'PLUS' 1
154               'END'
155             'FI'
156           'FI' ;
157         ORDER(ALPHAFILE[1:LSS],AUXORDER) ;
158         S 'MINUS' SCREENSDELETED ;
159           'C' SET LASTSETSIZE WHICH GIVES THE SCREEN-SET SIZE AT THE
160           END OF THIS ITERATION 'C'
161         LASTSETSIZE :=
162           'IF' LIMITKS > 0 'THEN' S := KS ; 'TO' KS 'WHILE'
163             TOTAL 'OF' ALPHAFILE[S] = LIMITKS 'DO' S 'PLUS' 1 ; S
164           'ELSE' S := KS ; 'WHILE' TOTAL 'OF' ALPHAFILE[S] = 0 'DO' S 'MINUS' 1;
165             S 'FI' ;
166         S := 0 ;
167         ORDER(ALPHAFILE[1:LASTSETSIZE],ALPHAORDER)
168       'END' ;
169
170     'PROC' RELATIVE ENTROPY = 'REAL' :
171       'BEGIN'
172         'INT' N := 0 ;
173         'FOR' X 'TO' LASTSETSIZE 'DO'
174           'BEGIN'
```

```
175             'REF'!KEY' AX = ALPHAFILE[X] ;
176             'IF' SIZE 'OF' AX < MAXLENGTH 'THEN'
177               'REF'!INT' T = NONREDUNDANT 'OF' AX ;
178               'INT' SAX = SIZE 'OF' AX + 1 ;
179               'INT' FIRST = (TYPE 'OF' AX)[1] ;
180               'BOOL' B := 'TRUE' ;
181               'FOR' Y 'FROM' X+1 'TO' LASTSETSIZE 'WHILE' B 'DO'
182                 'BEGIN'
183                   'REF'!KEY' AY = ALPHAFILE[Y] ;
184                   'IF' SIZE 'OF' AY = SAX 'THEN'
185                     'IF' (TYPE 'OF' AX)[1:SIZE 'OF' AX] 'EQUALS'
186                       (TYPE 'OF' AY)[1:SIZE 'OF' AX] 'THEN' T 'MINUS'
187                       TOTAL 'OF' AY 'ELSE' B := 'FALSE'
188                     'FI'
189                   'FI'
190                 'END'
191             'FI'
192           'END' ;
193         'FOR' X 'TO' LASTSETSIZE 'DO'
194           'IF' SIZE 'OF' ALPHAFILE[X] = 1 'THEN'
195           N 'PLUS' TOTAL 'OF' ALPHAFILE[X] 'FI' ;
196         'REAL' J := 0.0 ;
197         'FOR' X 'TO' LASTSETSIZE 'DO'
198           'BEGIN'
199             'INT' TKX = NONREDUNDANT 'OF' ALPHAFILE[X] ;
200             J 'PLUS'
201               'IF' TKX = 0 'THEN' 0.0
202               'ELSE'!'REAL' M = TKX/TOTALINCIDENCE ; M*LN(M) 'FI'
203           'END' ;
204           'C' Z IS THE NUMBER OF UNASSIGNED FRAGMENTS I . E. THE FRAGMENTS
205           FOR WHICH NO BITS WOULD BE SET USING THE CURRENT SCREEN SET 'C'
206         'INT' Z = TOTALINCIDENCE - N ;
207         -(J 'PLUS'!'IF' Z = 0 'THEN' 0.0
208             'ELSE'!'REAL' M = Z/TOTALINCIDENCE ; M*LN(M)'FI')/LN(LASTSETSIZE+1)
209       'END' ;
```

```
210
211     'PROC' READSCREEN = 'VOID' :
212       'BEGIN'
213         MTRD(1,RECORD[1]) ;
214         'FOR' X 'TO' MAXLENGTH 'DO'
215           (TYPE 'OF' CURRENT)[X] := 'ABS' RECORD[X+1] ;
216         SIZE 'OF' CURRENT := 'ABS' RECORD[MAXLENGTH+2] ;
217         TOTAL 'OF' CURRENT := NONREDUNDANT 'OF' CURRENT :=
218           'ABS' RECORD[MAXLENGTH+3]
219       'END' ;
220
221     'PROC' TRUNCATE SCREEN SET = 'VOID' :
222       'BEGIN'
223         'C' TRUNCATES THE SCREEN SET BY DELETING STRINGS IN SIZE AND
224         INVERSE FREQUENCY ORDER 'C'
225         'INT' X := LASTSETSIZE ;
226         'INT' SAFLS = MAXLENGTHPRESENT ;
227         ORDER(ALPHAFILE[1:LASTSETSIZE],SIZEORDER) ;
228         'WHILE' SIZE 'OF' ALPHAFILE[X] = SAFLS 'DO' X 'MINUS' 1 ;
229         ORDER(ALPHAFILE[X 'PLUS' 1:LASTSETSIZE],AUXORDER) ;
230         LASTSETSIZE := S := KS ;
231         ORDER(ALPHAFILE[1:S],ALPHAORDER) ;
232         'SKIP'
233       'END' ;
234
235     PRINT(("DESIRED SCREEN SET SIZE = ",KS+1,NEWLINE,NEWLINE,
236       "CALCULATED THRESHOLD FREQUENCY = ",THRESHOLD,NEWLINE)) ;
237     'WHILE' (READSCREEN ; SIZE 'OF' CURRENT = 1) 'DO'
238       'IF' (TOTAL 'OF' CURRENT > THRESHOLD)
239       'AND' (S < KS) 'THEN'
240         ALPHAFILE[S 'PLUS' 1] := CURRENT
241       'FI' ;
242     LASTSETSIZE := S ; S := 0 ;
243     'FOR' X 'FROM' LASTSETSIZE + 1 'TO' KS 'DO' TOTAL 'OF' ALPHAFILE[X] := 0 ;
244     ORDER(ALPHAFILE[1:LASTSETSIZE],ALPHAORDER) ;
```

```
245       'FOR' X 'FROM' 2 'TO' MAXLENGTH 'DO''BEGIN'
246         'WHILE' (READSCREEN ; SIZE 'OF' CURRENT = X) 'DO'
247           'IF' (S < KS)
248           'AND' (TOTAL 'OF' CURRENT > THRESHOLD) 'THEN'
249               FINDPARENT(X-1)
250           'FI' ;
251         UPDATEALPHAFILE(X=1)
252       'END' ;
253       'IF' LASTSETSIZE > KS 'THEN' TRUNCATE SCREEN SET
254       'ELSF' LASTSETSIZE < KS 'THEN' FAULT("INCOMPLETE SCREENSET") ; FREE
255       'ELSE' S := LASTSETSIZE 'FI' ;
256
257           'C' THE FIRST LINE OF THE OUTPUT SCREEN SET CONTAINS (1) THE MAXIMUM
258           POSSIBLE SCREEN LENGTH, (2) THE MAXIMUM ACTUAL SCREEN LENGTH AND
259           (3) THE NUMBER OF DIFFERENT ONE-INTS 'C'
260       PUT(SCREENSOUT,MAXLENGTH) ;
261       MAXLENGTH := MAXLENGTHPRESENT ;
262       'REAL' M := RELATIVE ENTROPY ;
263       PUT(SCREENSOUT,(S,MAXLENGTH)) ;
264       MAXLENGTH := 0 ;
265           'C' WE NOW DETERMINE THE NUMBER OF DIFFERENT ONE-INTS = THIS WILL
266           BE THE SIZE OF THE INDEX THAT IS REQUIRED TO SEARCH THE SCREEN SET 'C'
267       'FOR' X 'TO' S 'DO'
268         'IF' SIZE 'OF' ALPHAFILE[X] = 1 'THEN' MAXLENGTH 'PLUS' 1 'FI' ;
269       PUT(SCREENSOUT,MAXLENGTH) ;
270       PUT(SCREENSOUT,(TYPE 'OF' ALPHAFILE[S])[1]) ; 'C' WRITE OUT THE FIRST
271         INTEGER OF THE FINAL INTEGER STRING WHICH IS USED IN THE ASSIGNMENT
272         PROGRAM 'C'
273       'FOR' X 'TO' S 'DO' PUT(SCREENSOUT,(NEWLINE,SIZE 'OF' ALPHAFILE[X],
274                                  TYPE 'OF' ALPHAFILE[X])) ;
275       NEXTLINE(SCREENSOUT) ;
276       PRINT((NEWLINE,NEWLINE,"FINAL RELATIVE ENTROPY = ",M)) ;
277       MAXLENGTH := 0 ; 'FOR' X 'TO' S 'DO' MAXLENGTH 'PLUS'
278         NONREDUNDANT 'OF' ALPHAFILE[X] ;
279       PRINT((NEWLINE,NEWLINE,"TOTAL ASSIGNMENT FREQUENCY = ",MAXLENGTH)) ;
```

```
280       PRINT((NEWLINE,NEWLINE,"UNASSIGNED FREQUENCY = ", TOTALINCIDENCE -
281         MAXLENGTH)) ;
282       'TO' 15 'DO' NEXTLINE(STANDOUT) ;
283  FINIS :
284       MTEND(1,"CLOSE") ;
285       'SKIP'
286  'END'
287  'FINISH'
288  ****
```

APPENDIX II.

```
0  STRUCTURE MATCHING PROGRAM 'WITH' MTH 'FROM' :LIB.SHEF68ALBUM
1  'BEGIN'
2
3
4            'C' AUTHOR PETER WILLETT , PGSLIS , 1977 'C'
5            'C' THIS PROGRAM CARRIES OUT AN AUTOMATIC ANALYSIS OF A CHEMICAL
6            REACTION USING CONNECTION TABLE STRUCTURE REPRESENTATIVES.  IT IS
7            IN THREE MAIN PARTS.  THE FIRST ONE TAKES A "CROSSBOW" CONNECTION
8            TABLE AND CONVERTS IT TO A REDUNDANT CONNECTIVITY MATRIX.  THE
9            REACTANT AND PRODUCT MATRICES ARE THEN COMPARED, USING AN APPROXIMATE
10           STRUCTURE MATCHING PROCEDURE BASED ON THE MORGAN ALGORITHM, TO
11           IDENTIFY CERTAIN AREAS COMMON TO THE TWO STRUCTURES.  THE DIFFERENCES
12           ENGENDERED BY THE CHANGE ARE OBTAINED BY SUBTRACTION.  IN THE FINAL
13           SECTION, A VARIETY OF BITSCREENS ARE ASSIGNED TO CHARACTERISE BOTH
14           THE MOLECULES INVOLVED IN THE REACTION AND THE ANALYSIS ITSELF 'C'
15
16
17      'CHARPUT' ANALYSISFAILURE ; OPENC(ANALYSISFAILURE,LINE PRINTER,2) ;
18      'CHARPUT' MATRIXOUT ; OPENC(MATRIXOUT,LINE PRINTER,1) ;
19      [1:100]'INT' RPERCENT , PPERCENT ;
20      'CLEAR' RPERCENT ; 'CLEAR' PPERCENT ;
21      'MTFDA' FDAINPUT ; 'CLEAR' FDAINPUT ;
22      NAME 'OF' FDAINPUT := "ISICONNTABLE" ;
23      FGN 'OF' FDAINPUT := 1 ;
24      'MTFDA' FDAOUTPUT ; 'CLEAR' FDAOUTPUT ;
25      NAME 'OF' FDAOUTPUT := "REACTIONFILE" ;
26      FGN 'OF' FDAOUTPUT := 1 ;
27
28       'PROC' EXCEP := 'VOID' :
29        'BEGIN'
30          'IF' E 'OF' MTHWORD11 = 2
31           'THEN' PRINT((NEWLINE,"**END OF INPUT TAPE**" , DBC 'OF'
32             FDAINPUT)) ; 'GOTO' FINIS
33           'ELSE' PRINT((NEWLINE,"EXCEPTION",MTHWORD11)) ;
34             FAULT("PROGRAM TERMINATED")
```

```
      'FI'
    'END' ;

  MTDEF(1,0,0,FDAINPUT,EXCEP,1) ;
  MTDEF(2,1,512,FDAOUTPUT,EXCEP,2) ;
  'INT' REACTIONS := 0 ;
  'INT' GENERATED ADJACENCY TABLE := 0 ;
  'INT' OVERFLOWINMATCH := 0 ;
  'INT' SUCCESS := 0 ;
  'INT' NOMATCH := 0 , ALLMATCH := 0 ;

  'MODE''SCREEN' = 'STRUCT'('INT' SIZE,[1:6]'INT' TYPE) ;
  'MODE''INDEX' = 'STRUCT'('INT' TYPE,MAXSIZE,FIRST,LAST) ;
  'CHARPUT' BONDFILE,RINGFILE,ATOMFILE,ANALBONDFILE,ANALRINGFILE,
    ANALATOMFILE ;
  []'REF''CHARPUT' CP = (BONDFILE,RINGFILE,ATOMFILE,ANALBONDFILE,ANALRINGFILE,
    ANALATOMFILE) ;
  [1:6]'INT' INDEXLIMIT , SCREENSETSIZE , MAXPOSSIBLESCREENLENGTH ,
    MAXACTUALSCREENLENGTH ;
  'FOR' X 'TO' 6 'DO' OPENC(CP[X] , FILE READER , X) ;
  'FOR' X 'TO' 6 'DO' GET(CP[X],(MAXPOSSIBLESCREENLENGTH[X] , SCREENSETSIZE[X]
    ,MAXACTUALSCREENLENGTH[X],INDEXLIMIT[X])) ;
  [1:SCREENSETSIZE[1]]'SCREEN' BONDSCREEN ;
  [1:SCREENSETSIZE[2]]'SCREEN' RINGSCREEN ;
  [1:SCREENSETSIZE[3]]'SCREEN' ATOMSCREEN ;
  [1:SCREENSETSIZE[4]]'SCREEN' ANALBONDSCREEN ;
  [1:SCREENSETSIZE[5]]'SCREEN' ANALRINGSCREEN ;
  [1:SCREENSETSIZE[6]]'SCREEN' ANALATOMSCREEN ;
  [1:INDEXLIMIT[1]]'INDEX' BONDINDEX ;
  [1:INDEXLIMIT[2]]'INDEX' RINGINDEX ;
  [1:INDEXLIMIT[3]]'INDEX' ATOMINDEX ;
  [1:INDEXLIMIT[4]]'INDEX' ANALBONDINDEX ;
  [1:INDEXLIMIT[5]]'INDEX' ANALRINGINDEX ;
  [1:INDEXLIMIT[6]]'INDEX' ANALATOMINDEX ;
  [1:SCREENSETSIZE[1]+1,1:6]'INT' BONDPARENT ;
```

```
70      [1:SCREENSETSIZE[2]+1,1:3]'INT' RINGPARENT ;
71      [1:SCREENSETSIZE[3]+1,1:5]'INT' ATOMPARENT ;
72      [1:SCREENSETSIZE[4]+1,1:6]'INT' ANALBONDPARENT ;
73      [1:SCREENSETSIZE[5]+1,1:3]'INT' ANALRINGPARENT ;
74      [1:SCREENSETSIZE[6]+1,1:5]'INT' ANALATOMPARENT ;
75      []'REF'[]'INDEX' INDEXREF = (BONDINDEX,RINGINDEX,ATOMINDEX,ANALBONDINDEX,
76        ANALRINGINDEX,ANALATOMINDEX) ;
77      []'REF'[]'SCREEN' SCREENREF = (BONDSCREEN,RINGSCREEN,ATOMSCREEN,
78        ANALBONDSCREEN,ANALRINGSCREEN,ANALATOMSCREEN) ;
79      []'REF'[,]'INT' PARENTREF = (BONDPARENT,RINGPARENT,ATOMPARENT,
80        ANALBONDPARENT,ANALRINGPARENT,ANALATOMPARENT) ;
81      'CLEAR' BONDPARENT ; 'CLEAR' RINGPARENT ; 'CLEAR' ATOMPARENT ;
82      'CLEAR' ANALBONDPARENT ; 'CLEAR' ANALRINGPARENT ; 'CLEAR' ANALATOMPARENT ;
83
84      'STRUCT'([1:1]'BYTES' LENGTH,[1:1]'BITS' MINIMUMREQUIREMENTS,
85        [1:4]'BITS' MOLFORMBITS,
86        [1:4,0:1]'BITS' RINGBITS,[1:2,0:9]'BITS' MOLATOMBITS,[1:2,0:9]'BITS'
87        MOLBONDBITS,[1:2,0:9]'BITS' ANALATOMBITS,[1:2,0:9]'BITS' ANALBONDBITS,
88        [1:400]'CHAR' WLNFRAGMENTS,[1:200]'CHAR' WLN,[1:20]'CHAR'
89        BIBDETAILS)OUTPUTBUFFER ;
90      'REF''BITS' MINIMUM = (MINIMUMREQUIREMENTS 'OF' OUTPUTBUFFER)[1] ;
91      'REF'[]'CHAR' RWLN = (WLN 'OF' OUTPUTBUFFER)[1:100] , PWLN =
92        (WLN 'OF' OUTPUTBUFFER)[101:200] ;
93      (LENGTH 'OF' OUTPUTBUFFER)[1] := "003Y" ;
94      'STRUCT' ([1:1]'BYTES' LENGTH , [1:2300]'CHAR' CROSSBOW)BUFFER ;
95      'REF'[]'CHAR' C = CROSSBOW 'OF' BUFFER ;
96      'INT' PA ; 'C' POINTER TO THE CROSSBOW RECORD 'C'
97
98      [1:100]'INT' RCONNECTIVITY , PCONNECTIVITY ;
99      [1:150,1:2]'INT' RBONDTABLE , PBONDTABLE ;
100     [1:9]'INT' RMOLFORM , PMOLFORM ;
101     'INT' RBONDCOUNT , PBONDCOUNT ;
102     [1:100]'CHAR' RATOMLIST , PATOMLIST ;
103     'INT' RATOMCOUNT , PATOMCOUNT ;
104     [1:100,1:4]'INT' RADJACENCY , PADJACENCY ;
```

```
[1:100]'BOOL' RDELETED , PDELETED ;
[1:100]'BOOL' RRINGATOMTEST , PRINGATOMTEST ;
[1:100]'CHAR' RUNITS ; PUNITS ;
'BOOL' REACTANT ;

'OP''TWOFOUR' = ('REF''INT' X)'INT' :
  'BEGIN'
    'INT' Z := X '/:=' 24 ;
    'IF' Z = 0 'THEN' Z := 24 ; X 'MINUS' 1 'FI' ;
    Z
  'END' ;

'OP''EQUALS' = ('REF'[]'INT' A,B)'BOOL' :
  'BEGIN'
    'INT' X := 0 ; 'BOOL' SAME := 'TRUE' ;
    'TO''UPB' A 'WHILE' SAME 'DO' (X 'PLUS' 1 ; SAME := A[X] = B[X] ) ;
    SAME
  'END' ;

'OP''GREATERTHAN' = ('REF'[]'INT' A,B)'BOOL' :
  'BEGIN'
    'INT' X := 0 ; 'BOOL' SAME := 'TRUE' ;
    'TO''UPB' A 'WHILE' SAME 'DO' (X 'PLUS' 1 ; SAME := A[X] = B[X] ) ;
    'IF' A[X] > B[X] 'THEN''TRUE''ELSE''FALSE''FI'
  'END' ;

'OP''LESSTHAN' = ('REF'[]'INT' A,B)'BOOL' :
  'BEGIN'
    'INT' X := 0 ; 'BOOL' SAME := 'TRUE' ;
    'TO''UPB' A 'WHILE' SAME 'DO' (X 'PLUS' 1 ; SAME := A[X] = B[X] ) ;
    'IF' A[X] < B[X] 'THEN''TRUE''ELSE''FALSE''FI'
  'END' ;

'FOR' X 'TO' 6 'DO''BEGIN'
      'C' READS IN A SCREEN SET AND SETS UP THE INDEX, IE A LIST OF
```

```
140           THE ONE-INTS TOGETHER WITH THE ASSOCIATED MAXIMUM LENGTH POSSIBLE
141           AND THE RANGE OF SCREEN NUMBERS FOR WHICH THAT IS THE FIRST INTEGER
142           (THIS DATA IS USED TO LIMIT THE EXTENT OF THE BINARY SEARCH)'C'
143         'REF''CHARPUT' CPX = CP[X] ; 'REF'[]'SCREEN' SCREEN = SCREENREF[X] ;
144         'REF'[]'INDEX' INDEX = INDEXREF[X] ; 'INT' MAXSCREENSIZE =
145         MAXPOSSIBLESCREENLENGTH[X] ; 'REF'[,]'INT' PARENT = PARENTREF[X] ;
146         'INT' LENGTH := 1 , NUMBER := 1 ;
147         'INT' ONEINTS := 0 ;
148         GET(CPX,(NEWLINE,SIZE 'OF' SCREEN[1],(TYPE 'OF' SCREEN[1])[1:
149           MAXSCREENSIZE])) ;
150         'INT' SSS = 'UPB' SCREEN ;
151         'FOR' X 'FROM' 2 'TO' SSS 'DO'
152           'BEGIN'
153             'REF''SCREEN' SX = SCREEN[X] , SX1 = SCREEN[X-1] ;
154             'INT' J = (TYPE 'OF' SX1)[1] ;
155             GET(CPX,(NEWLINE,SIZE 'OF' SX,(TYPE 'OF' SX)[1:MAXSCREENSIZE])) ;
156             'C' WE SCAN BACK THROUGH THE SCREENS TO DATE TO DETECT THE
157             ANTECEDENTS OF THE CURRENT SCREEN; THESE MORE GENERAL SCREENS
158             WILL BE ALLOCATED AT THE SAME TIME AS THE CURRENT SCREEN TO ALLOW
159             FOR MORE GENERIC SEARCHES 'C'
160             'REF''INT' SSX = SIZE 'OF' SX ;
161             'IF' SSX > 1 'THEN'
162               [1:SSX]'INT' COPY := (TYPE 'OF' SX)[1:SSX] ; COPY[SSX] := 0 ;
163               'REF'[]'INT' PX = PARENT[X] ;
164               'INT' Y := 1 ;
165               'FOR' Z 'FROM' X-1 'BY' -1 'TO' 1 'WHILE' Y < SSX 'DO'
166                 'IF' (TYPE 'OF' SCREEN[Z])[1:SSX] 'EQUALS' COPY 'THEN'
167                   COPY[SSX-Y] := 0 ; PX[Y 'PLUS' 1] := Z 'FI' ;
168               PX[1] := Y-1
169             'FI' ;
170             'IF' J # (TYPE 'OF' SX)[1] 'THEN'
171             INDEX[ONEINTS 'PLUS' 1] := ((TYPE 'OF' SX1)[1],LENGTH,NUMBER,X) ;
172               NUMBER := X ; LENGTH := 1
173             'ELSF' SIZE 'OF' SX > LENGTH 'THEN' LENGTH := SIZE 'OF' SX 'FI'
174           'END' ;
```

```
175         INDEX[ONEINTS 'PLUS' 1] := ((TYPE 'OF' SCREEN[SSS])[1],LENGTH,
176                                     NUMBER,SSS+1)
177       'END' ;
178
179     'PROC' SEARCH = ('REF'[]'INT' QUERY STRING,'REF'[]'INDEX' INDEX,
180       'REF'[,]'INT' GENERIC,'REF'[]'SCREEN' SCREEN,'REF'[]'BITS' BITS) :
181       'BEGIN'
182           'C' THIS SCREEN LOOK-UP PROCEDURE WORKS IN 2 STAGES.  FIRSTLY,
183           THE SCREEN-SET INDEX IS SEARCHED TO FIND THE LONGEST POSSIBLE
184           SCREEN GIVEN THE FIRST INTEGER OF THE QUERY STRING.  IF THIS LENGTH
185           IS GREATER THAN 1 WE THEN SEARCH THE SCREEN-SET ITSELF USING INFORMAT-
186           ION IN THE INDEX ENTRY TO LIMIT THE WIDTH OF THE SEARCH.  IF THE SCREE
187           IS NOT FOUND IN THE INDEX, THE NULL SCREEN IS AUTOMATICALLY ALLOCATED
188           WITHOUT ANY FURTHUR ACTION 'C'
189         'INT' LOWER := 0 ; UPPER := 'UPB' INDEX + 1 ,
190           QUERY := QUERY STRING[1] , FOUND := 0 , MIDPOINT :=
191           'ENTIER'((LOWER+UPPER)/2) ;
192         'WHILE' MIDPOINT # LOWER 'DO'
193           'BEGIN'
194             'INT' TJ = TYPE 'OF' INDEX[MIDPOINT] :
195             'IF' QUERY = TJ 'THEN' FOUND := LOWER := MIDPOINT
196             'ELSF' QUERY > TJ 'THEN' LOWER := MIDPOINT :
197               MIDPOINT := 'ENTIER'((LOWER+UPPER)/2)
198             'ELSE' UPPER := MIDPOINT ;
199               MIDPOINT := 'ENTIER'((LOWER+UPPER)/2) 'FI'
200           'END' ;
201         FOUND :=
202           'IF' FOUND > 0 'THEN' 'REF''INDEX' IJ = INDEX[FOUND] :
203             'IF' MAXSIZE 'OF' IJ > 1 'THEN'
204               'BEGIN'
205                 'C' SEARCH MAIN SCREEN SETUSING LIMITS FROM THE INDEX ENTRY'C'
206                 'INT' FOUND := 0 ;
207                 'FOR' X 'FROM' MAXSIZE 'OF' IJ 'BY' -1 'TO' 2 'WHILE' FOUND = 0
208                   'DO''BEGIN'
209                     'REF'[]'INT' AX = QUERY STRING[2:X] :
```

```
                    'INT' L := FIRST 'OF' IJ - 1 , R := LAST 'OF' IJ ,
                        J := 'ENTIER'((L+R)/2) :
                    'WHILE' J # L 'DO''BEGIN'
                        'REF''SCREEN' SJ = SCREEN[J] ;
                        'REF'[]'INT' SJX = (TYPE 'OF' SJ)[2:X] :
                        'IF' (SIZE 'OF' SJ = X)'AND'(AX 'EQUALS' SJX) 'THEN'
                          FOUND := L := J
                        'ELSE' SJX 'LESSTHAN' AX 'THEN' L := J : J := 'ENTIER'
                          ((L+R)/2)
                        'ELSE' R := J ; J := 'ENTIER'((L+R)/2) 'FI'
                      'END'
                  'END' ;
                'C' ALLOCATE THE ONE-INT SCREEN IF A LARGER ONE HAS NOT
                BEEN TRACED IN THE BINARY SEARCH 'C'
                'IF' FOUND # 0 'THEN' FOUND 'ELSE' FIRST 'OF' IJ 'FI'
              'END'
            'ELSE' FIRST 'OF' IJ 'FI'
          'ELSE''UPB' SCREEN + 1 'FI' ;
        'PROC' SETBIT = ('REF''INT' X) : 'BEGIN''INT' M = 'TWOFOUR' X :
          'REF''BITS' SF = BITS[X] ; SF := M 'SET' SF
        'END' ;
        'C' WE NOW SET THE BIT FOR THE CURRENT SCREEN AND FOR ITS PARENTS 'C'
      'REF'[]'INT' GF = GENERIC[FOUND] ; SETBIT(FOUND) ;
      'FOR' X 'TO' GF[1] 'DO' ('INT' I := GF[X+1] ; SETBIT(I))
    'END' ;


'PROC' CREATE ADJACENCY TABLE = ('REF'[,]'INT' ADJACENCY,BOND TABLE,
  'REF'[]'INT' CONNECTIVITY,MOLFORM,'REF''INT' ATOMCOUNT,BONDCOUNT,
  'REF'[]'BOOL' RINGATOMTEST,'REF'[]'CHAR' WLN,UNITS,ATOMLIST)'BOOL' :
  'BEGIN'

        'C'
        THIS PROGRAM CONVERTS A CROSSBOW TABLE TO A REDUNDANT ATOM-BOND
        CONNECTION TABLE.  THE CROSSBOW DATA HAS BEEN
```

```
245             STORED IN THE AREA "C", THE EXACT POSITION
246             WITHIN THIS AT ANY TIME BEING GIVEN BY THE VALUE OF "P"
247             "PA" IS USED TO MARK THE DEMARCATION POINTS BETWEEN THE VARIOUS
248             SECTIONS OF THE CROSSBOW RECORD.  THE SECTIONS ARE
249             (1) UNITS : 150 CHARACTERS
250             (2) CONNECTION TRANSFERS : 40 3-DIGIT NUMBERS
251             (3) RINGS : 27 4-CHARACTER STRINGS
252             (4) RING CONNECTIONS : 20 3-DIGIT NUMBERS
253             (5) MODIFIERS : 91 CHARACTERS
254             (6) WLN : 102 CHARACTERS
255             (7) RINGSCREEN : 150 CHATACTERS
256             'C'
257
258         'INT' P := PA ;
259         'INT' FAIL := 99 ;
260         'STRING' FAILURE ;
261         'BOOL' OK := 'TRUE' ;
262         'INT' ATOMNUM ;
263         [1:40]'INT' CONNTRAN ;
264         [1:20]'INT' RINGCONN ;
265         'CLEAR' CONNTRAN , 'CLEAR' RINGCONN ;
266         'BOOL' ACYCLIC := 'FALSE' , B := 'TRUE' ;
267         'REF'[]'BITS' RSBIT = 'IF' REACTANT 'THEN' (RINGBITS 'OF'
268           OUTPUTBUFFER)[1] 'ELSE' (RINGBITS 'OF' OUTPUTBUFFER)[3] 'FI' ;
269
270         'PROC' STRANGEATOM = 'VOID' :
271           (OK := 'FALSE' ; 'GOTO' EXIT) ;
272
273             'C' INPUT INTEGERS ARE WRITTEN ON TAPE AS 3-DIGIT CHARACTERS
274             AND THIS PROCEDURE CONVERTS THEM TO INTEGER FORMAT FOR SUBSEQUENT
275             MANUPULATION 'C'
276         'PROC' INTEGER = 'INT' :
277           'BEGIN'
278             'INT' VALUE := 100*'ABS'C[P] + 10*'ABS' C[P+1] + 'ABS' C[P+2] :
279             P 'PLUS' 3 ;
```

```
        VALUE
      'END' ;

          'C' SET UP UNITS SECTION IE "DOT-PLOT" SYMBOLS 'C'
    'WHILE' C[P] # " " 'DO' P 'PLUS' 1 ;
    'IF' P = 1 'THEN' FAILURE := "MISSING RECORD" ; STRANGEATOM 'FI' ;
          'C' ONLY MOLECULES CONTAINING < 100 UNITS ARE PERMITTED 'C'
    ('INT' M = P - PA ;
     'IF' M > 100 'THEN' FAILURE := "TOO MANY ATOMS" ;
      STRANGEATOM 'ELSE' UNITS [1:M] := C[PA:P-1] 'FI' ) ;
          'C' READ IN CONNECTION TRANSFERS . THE FIRST NUMBER IS THE ATOM
          AT WHICH CONNECTIVITY IS BROKEN AND THE SUBSEQUENT ONE THAT AT
          WHICH IT IS RESUMED.  999 IS A DEMARCATION SYMBOL BEFORE
          THE FINAL 3-DIGIT STRING WHICH IS THE TOTAL NUMBER OF ATOMS
          IN THE MOLECULE 'C'
    P := PA 'PLUS' 150 ;
    'FOR' X 'TO' 40 'WHILE' B 'DO'
      'IF' C[P] = " " 'THEN' B := 'FALSE'
      'ELSE''REF''INT' CX = CONNTRAN[X] ; CX := INTEGER ;
        'IF' CX = 0 'THEN' FAILURE := "ZERO IN CONNTRAN" ; STRANGEATOM
        'ELSE' CX = 999 'THEN' ATOMNUM := INTEGER ; B := 'FALSE''FI'
      'FI' ;

    'IF' ATOMNUM > 96 'THEN' FAILURE := "TOO MANY ATOMS" ; STRANGEATOM 'FI' ;
    [1:ATOMNUM,1:ATOMNUM]'INT' CONNECTION TABLE ;
    'REF'[]'INT' CTONE = CONNECTION TABLE[,1] ;
    'CLEAR' CONNECTION TABLE ;

        'C' AS WE IGNORE THE MODIFIERS SECTION IN THE ICI PROGRAM , THIS
          PROGRAM ONLY HANDLES THE ATOMS CNOSP AND HAL. IN A LIMITED RANGE
          OF VALENCE STATES . ANY OTHER ATOM/VALENCE TYPES CAUSES THE FAIL
          ROUTINE "STRANGEATOM" TO BE CALLED 'C'
    'FOR' X 'TO' ATOMNUM 'DO'
      'BEGIN'
        'INT' AUX = 'ABS' UNITS[X] ;
```

```
      'IF' AUX < 12 'THEN' FAILURE := "STRANGE ATOM DETECTED" ; STRANGEATOM
        'ELSE''CHAR' J := ATOMLIST[X] := 'CASE' AUX-11 'IN'
          "C","*","N","*","*","*","*","N","*","P","P","*","*","*","*","*",
          "*","*","*","S","N","*","P","C","C","E","F","G","*","I","N","N",
          "C","N","N","O","P","O","*","S","C","C","*","S","C","C","N",
          "*","*","*","*","*" 'ESAC' ;
          'IF' J = "*" 'THEN' FAILURE := "STRANGE ATOM DETECTED" :
           STRANGEATOM 'FI'
      'FI'
    'END' ;

        'C' "EQUIVALENT" CONTAINS PAIRS OF DUPLICATE ATOMS (FUSION POINTS) :
        "MONOCYCLE" CONTAINS THE INITIAL AND FINAL ATOMS AND THE SIZE OF EACH
        MONOCYCLE IN THE MOLECULE 'C'
[1:25,1:2]'INT' EQUIVALENT ; 'CLEAR' EQUIVALENT ;
[1:25,1:3]'INT' MONOCYCLE ; 'CLEAR' MONOCYCLE ;
'REF'[]'INT'  E1 = EQUIVALENT[,1] , E2 = EQUIVALENT[,2] ,
              MC1 = MONOCYCLE[,1] , MC2= MONOCYCLE[,2] ;

'PROC' MAKEBOND = ('INT' A,B)'VOID' :
  (CONNECTION TABLE[A,B] := CONNECTION TABLE[B,A] := 1 ) ;

'PROC' ERASEBOND = ('INT' A,B)'VOID' :
  ('IF' A = 0 'OR' B = 0 'THEN' FAILURE := "ERASEBOND FAILURE" ;
    STRANGEATOM 'ELSE'
     CONNECTION TABLE[A,B] := CONNECTION TABLE[B,A] := 0
   'FI' ) ;

        'C' THE NEXT SECTION READS IN THE RINGS SECTION OF THE CROSSBOW
        RECORD ( SEE ICI PAMPHLET FOR FORMAT DETAILS ).  THIS ALLOWS THE
        IDENTIFICATION OF ALL MONOCYCLES AND RING SYSTEMS PRESENT IN THE
        MOLECULE 'C'
'INT' E := 0 ; 'C' E COUNTS THE NUMBER OF ATOMS THAT HAVE BEEN
                COUNTED TWICE I.E. FUSION ATOMS 'C'
'INT' RINGCOUNT := 0 ; 'C' COUNTS THE NUMBER OF MONOCYCLES PRESENT 'C'
```

```
'INT' X := 0 ;
[1:27]'INT' RSA , RSB ; 'C' THESE TWO ARRAYS CONTAIN THE NUMBERS OF
                         THE FIRST AND LAST ATOMS OF EACH RING SYSTEM IN THE
                         MOLECULE 'C'
'CLEAR' RSA ; 'CLEAR' RSB ;
'INT' RS := 0 ;
      'C' "RING SYSTEM" IS USED TO DETECT 2 4-DIGIT (ONE 3-DIGIT
      RINGSTART AND A SINGLE-DIGIT RINGSIZE) STRINGS IN SUCCESSION ; THIS
      MARKS THE END OF ONE RING SYSTEM AND THE START OF ANOTHER 'C'
'STRUCT' ('INT' LASTONE , ATOMCOUNT) RING SYSTEM := (1,999) ;
B := 'TRUE' ;
'INT' RINGSTART ;
P := PA 'PLUS' 120 ;
'TO' 27 'WHILE' B 'DO'
  'BEGIN'
    'CHAR' CP = C[P] ;
    X 'PLUS' 1 ;
    'IF' CP # " " 'THEN'
      'INT' V ; RINGSIZE ;
      'IF' 'ABS' CP <= 9 'THEN'
        'INT' ACP = 'ABS' C[P+3] ;
        RINGSTART := INTEGER ;
        P 'PLUS' 1 ; 'C' ADVANCE "P" TO START OF NEXT 4-CHAR STRING 'C'
        'IF' LASTONE 'OF' RING SYSTEM = X - 1 'THEN'
          RSB[RS 'PLUS' 1] := ATOMCOUNT 'OF' RING SYSTEM 'FI' ;
        RINGSIZE := 'IF' ACP = 0 'THEN' 6
                    'ELSF' ACP <= 9 'THEN' ACP
                    'ELSE' ACP - 24
                  'FI' ;
        V := RINGSIZE + RINGSTART - 1 ;
        RING SYSTEM := (X,V) ;
        MONOCYCLE[RINGCOUNT 'PLUS' 1] := (RINGSTART,V,RINGSIZE) ;
          'C' CREATE RING-CLOSURE BOND 'C'
        MAKEBOND(V,RINGSTART)
      'ELSF' CP = "P" 'OR' CP = "B"
```

```
385                 'THEN' FAILURE := "COMPLEX RING SYSTEM" ; STRANGEATOM
386               'ELSF' CP = "S" 'THEN'
387                 'C' SPIRO RING PAIR 'C'
388                 P 'PLUS' 1 ; E1[E 'PLUS' 1] := INTEGER ; P 'PLUS' 1 ;
389                 RSB[RS 'PLUS' 1] := ATOMCOUNT 'OF' RING SYSTEM ;
390                 E2[E] := INTEGER
391               'ELSE' P 'PLUS' 1 ; EQUIVALENT[E 'PLUS' 1] := (RINGSTART.INTEGER) ;
392                 P 'PLUS' 1 ; EQUIVALENT[E 'PLUS' 1] := (V,INTEGER)
393               'FI'
394             'ELSE' B := 'FALSE'
395             'FI'
396         'END' ;
397       'IF' LASTONE 'OF' RING SYSTEM = X - 1 'THEN'
398         RSB[RS 'PLUS' 1] := ATOMCOUNT 'OF' RING SYSTEM ;
399         RSA[RS] := RINGSTART 'FI' ;
400             'C' CHECK TO SEE WHETHER THE MOLECULE IS ACYCLIC 'C'
401       'IF' X = 1 'THEN' ACYCLIC := 'TRUE''FI' ;
402             'C' SET UP RING CONNECTIONS SECTION OF TABLE 'C'
403       P := PA 'PLUS' 108 ; B := 'TRUE' ;
404       'INT' RC := 0 ; 'C' POINTER TO POSITION IN RINGCONN 'C'
405       'TO' 20 'WHILE' C[P] # " " 'DO''BEGIN'
406           'REF''INT' RCRC = RINGCONN[RC 'PLUS' 1] ;
407           RCRC := INTEGER ;
408           'IF' RCRC > ATOMNUM 'THEN' FAILURE := "RCRC FAILURE" ; STRANGEATOM 'FI'
409         'END' ;
410
411             'C' KNOWING THE INITIAL AND FINAL ATOMS OF EACH MONOCYCLE, AND ALSO
412             THE ENTRY AND EXIT POINTS FROM EACH RING SYSTEM, WE CAN DETERMINE THE
413             INITIAL ATOMS IN EACH RING SYSTEM, AT THE SAME TIME CHECKING THAT
414             IF AN ENTRY IN "RSA" CORRESPONDS TO A DUPLICATE ATOM, THE LOWER
415             NUMBERED ALTERNATIVE IS CHOSEN 'C'
416       RS := 0 ;
417       'IF' RC > 2 'THEN''FOR' J 'BY' 2 'TO' RC-1 'DO''BEGIN'
418           'INT' RCJ = RINGCONN[J] , RCJ1 = RINGCONN[J+1] ;
419           'INT' M = 'IF' RCJ = 0 'THEN' RCJ1
```

```
                    'ELSE' RCJ1 = 0 'THEN' RCJ
                    'ELSE' RCJ > RCJ1 'THEN' RCJ1 'ELSE' RCJ 'FI' ;
        'INT' L := 0 ;
        'REF'[]'INT' MC = MONOCYCLE[,1] ;
        'BOOL' B := 'TRUE' ;
        'TO' RINGCOUNT 'WHILE' B 'DO''BEGIN'
            L 'PLUS' 1 ;
            'IF' M < MC[L] 'THEN' B := 'FALSE' ; L 'MINUS' 1
            'ELSE' M = MC[L] 'THEN' B := 'FALSE''FI'
          'END' ;
        RSA[RS 'PLUS' 1] := ('INT' MCL = MC[L] ; 'BOOL' B := 'FALSE' ; L := 0;
          'TO' E 'WHILE''NOT' B 'DO' B := MCL = E2[L 'PLUS' 1] ;
          'IF' B 'THEN' E1[L] 'ELSE' MCL 'FI' )
    'END'
'ELSE' RSA[RS 'PLUS' 1] := MC1[1] 'FI' ;

'PROC' RINGSCREENS = ('REF'[]'INT' MC) :
    'BEGIN'
        'C' SET UP RING SCREENS FOR REACTING MOLECULES ; THERE ARE 3 LEVELS
        OF DESCRIPTION :
        1 RING SIZE
        2 NUMBER OF HETEROATOMS + 50
        3 TYPE OF HETEROATOMS UNLESS THERE ARE NOT ANY IN WHICH CASE( IE A
        CARBOCYCLIC RING) THE RING IS DESCRIBED BY THE TOTAL NUMBER OF
        EXTRA-RING CONNECTIONS 'C'
      [1:3]'INT' RING ;
      'BYTES' NOSP = "NOSP" ;
      'INT' HET := 0 , NC := 0 , EXTERNAL := 0 ;
      'FOR' M 'FROM' MC[1] 'TO' MC[2] 'DO'
        'BEGIN'
          'CHAR' AM = ATOMLIST[M] ;
          'IF' AM = "C" 'THEN' EXTERNAL 'PLUS'
            'CASE'('ABS'UNITS[M]-51)'IN' 1,1,0,0,0,2,1 'OUT' 0 'ESAC'
          'ELSE''INT' Y := 1 ;
            'WHILE' AM # Y 'ELEM' NOSP 'DO' Y 'PLUS' 1 ;
```

```
455                 HET 'PLUS' 10↑(Y-1) ; NC 'PLUS' 1 'FI'
456               'END' ;
457             'IF' NC = 0 'THEN' HET := EXTERNAL 'FI' :
458             RING := (NC[3],NC+50,HET) ;
459             SEARCH(RING,RINGINDEX,RINGPARENT,RINGSCREEN,RSBIT)
460           'END' ;
461
462       'FOR' X 'TO' RINGCOUNT 'DO' RINGSCREENS(MONOCYCLE[X]) ;
463       ('INT' P = 'IF' REACTANT 'THEN' 1 'ELSE' 5 'FI' ;
464       'IF' RINGCOUNT > 0 'THEN''FOR' X 'FROM' P 'TO''IF' RINGCOUNT>2 'THEN' P+1
465         'ELSE' P 'FI''DO' MINIMUM := X 'SET' MINIMUM
466       'FI') ;
467
468             'C' PA IS ADVANCED TO THE START OF THE WLN STRING SKIPPING THE
469             MODIFIERS SECTION : THIS PROGRAM WILL ONLY HANDLE FULLY COVALENT
470             MOLECULES CONTAINING A LIMITED NUMBER OF ATOM TYPES IN THEIR NORMAL
471             VALENCE STATES 'C'
472       P := PA 'PLUS' 151 ;
473       B := 'TRUE' ;
474             'C' TWO CONSECUTIVE SPACES SIGNALS THE END OF THE WLN SRING 'C'
475       'TO' 102 'WHILE' B 'DO'
476         'IF' C[P] = " " 'AND' C[P 'PLUS' 1] = " " 'THEN' B := 'FALSE'
477         'ELSE' P 'PLUS' 1 'FI' ;
478       ('INT' M := P-PA ;
479        'IF' M > 100 'THEN' FAILURE := "TOO MANY ATOMS" ; STRANGEATOM
480       'ELSE' WLN[1:M] := C[PA:P-1] 'FI' ;
481           'C' THE ICI PROGRAM PRODUCES AN INCORRECT TABLE IF THE INITIAL WLN
482             SYMBOL IS "R" 'C'
483       'IF' WLN[1] = "R" 'THEN' FAILURE := "INITIAL BENZENE" : STRANGEATOM 'FI' :
484       P := PA 'PLUS' 102 ; 'C' ADVANCE TO START OF ICI RINGSCREEN 'C'
485       M := 0 ; 'FOR' X 'FROM' P 'TO' P+ATOMNUM-1 'DO''BEGIN'
486           'CHAR' CX = C[X] ; RINGATOMTEST[M 'PLUS' 1] :=
487             'IF' CX = "0" 'OR' CX = "D" 'THEN''FALSE''ELSE''TRUE''FI'
488         'END' ) ;
489
```

```
490       'PROC' FORM CONNECTION TABLE ENTRIES = 'VOID' :
491         'BEGIN'
492             'C'
493             THE CONNECTION TRANSFERS AND RING CONNECTORS INFORMATION ARE
494             USED TO DETERMINE THE ATOM AT THE TOP OF THE BONDING-STACK -
495             "LASTATOM" - AND BONDS ARE THEN MADE BETWEEN THIS AND THE CURRENT
496             ATOM DENOTED BY THE SUBSCRIPT "M".
497             TWO EXCEPTIONS ARE MADE TO THIS RULE
498             (1) IF "LASTATOM" IS THE EXIT POINT FROM ONE RING SYSTEM AND "M"
499             CORRESPONDS TO THE FIRST MEMBER OF A SECOND
500             (2)IF "LASTATOM" IS THE FINAL ATOM IN ONE MONOCYCLE IN A RING SYSTEM
501             OF MORE THAN 2 RINGS AND "M" THE FIRST IN A SECOND MONOCYCLE.
502             'C'
503           'INT' DONE := 1 ; LASTATOM := 1 , CT := 1 , RC := 1 ;
504           RS := 1 ;
505           'FOR' M 'FROM' 2 'TO' ATOMNUM 'DO'
506             'BEGIN'
507               'REF''BOOL' IN RING = RINGATOMTEST[M] ;
508               'IF' M = MC2['DONE] 'THEN' DONE 'PLUS' 1 'FI' ;
509               MAKEBOND(M,LASTATOM) ;
510               'IF' IN RING 'AND' RC > 1 'AND' (LASTATOM=RINGCONN[RC-1] 'OR'
511                 M=RSA[RS]) 'THEN' ERASEBOND(M,LASTATOM) ; MAKEBOND(RINGCONN[RC],
512                 LASTATOM) 'FI' ;
513               'IF' IN RING 'AND' DONE > 1 'AND' LASTATOM = MC2['DONE-1] 'AND'
514                 M # RSA[RS] 'AND'
515                 ('INT' J=MC1['DONE-1] ; 'INT' Z := 1 ; 'TO' E-1 'WHILE' J # E1[Z]
516                 'DO' Z 'PLUS' 1 ; M # E2[Z]) 'THEN' ERASEBOND(M,LASTATOM) 'FI' ;
517               LASTATOM :=
518                 'IF' M = RSB[RS] 'THEN'
519                   'INT' J = RINGCONN[(RC 'PLUS' 2)-1] ; RS 'PLUS' 1 ;
520                   'IF' J # 0 'THEN' J 'C' IE THERE ARE SEPARATE ENTRY AND EXIT
521                                       POINTS FOR THE RING SYSTEM 'C'
522                   'ELSF' M # CONNTRAN[CT] 'THEN' M 'ELSE'
523                     CONNTRAN[(CT 'PLUS' 2)-1] 'FI'
524                 'ELSF' M # CONNTRAN[CT] 'THEN' M 'ELSE' CONNTRAN[(CT 'PLUS' 2)-1]
```

```
525                 'FI'
526             'END'
527         'END' ;
528
529
530     'PROC' NOTE DUPLICATE ATOMS = ('INT' J) :
531       'BEGIN'
532           'C'
533             THE CONNECTION TABLE IS SCANNED ROW BY ROW FOR NON-ZERO ENTRIES
534             WHILST DUPLICATE ATOMS ARE NOTED BY ENTERING A BOND ORDER
535             OF 9 AT THE START OF THE APPROPRIATE ROW AND COLUMN OF THE ARRAY
536             "CONNECTION TABLE" AND THEIR CONNECTIONS ARE TRANSFERRED TO THEIR
537             EQUIVALENT ATOMS.
538             'C'
539           'REF'[]'INT' M = EQUIVALENT[J] ;
540           'INT' A := M[1] ; B := M[2] ;
541           'FOR' BATOMS 'TO' ATOMNUM 'DO'
542             'BEGIN'
543               'REF'[]'INT' CT1 = CONNECTION TABLE[BATOMS] , CT2 = CONNECTION
544                 TABLE[,BATOMS] ;
545               'IF' CT2[B] # 0 'THEN' CT1[A] := CT2[A] := 1 ; CT1[B] := 0 'FI'
546             'END' ;
547           CTONE[B] := 9
548       'END' ;
549
550     FORM CONNECTION TABLE ENTRIES ;
551     'FOR' X 'FROM' 6 'BY' -1 'TO' 1 'DO' NOTE DUPLICATE ATOMS(X) :
552     'FOR' X 'TO' ATOMNUM 'DO' CONNECTION TABLE[X,X] := 0 ;
553
554           'C'
555             A REDUNDANT ADJACENCY LIST IS BUILT UP FROM THE CONNECTION TABLE
556             AVOIDING THE DUPLICATE ATOMS . AT THE SAME TIME "BOND TABLE" IS
557             CREATED WHICH LISTS THE CONSTITUENT ATOMS OF ALL BONDS IN THE
558             MOLECULE NON-REDUNDANTLY .
559             THE ADJACENCY TABLE CONTAINS A 4-ELEMENT VECTOR, PREVIOUSLY ZERO
```

```
560           FILLED, FOR EACH ATOM IN THE STRUCTURE: THE NON-ZERO ELEMENTS CONTAIN
561           THE NUMBERS OF THE ATTACHED ATOMS.  THE BOND TABLE CONTAINS TWO
562           ELEMENTS FOR EACH BOND PRESENT AND THEY ARE FILLED BY THE NUMBERS
563           OF THE CONSTITUENT ATOMS 'C'
564       'INT' DUPSDONE := 0 ;
565       BONDCOUNT := 0 ;
566       'FOR' ROW 'TO' ATOMNUM 'DO'
567         'BEGIN'
568         'REF'[]'INT' CTROW = CONNECTION TABLE[ROW, ] ;
569         'IF' CTROW[1] # 9 'THEN''INT' ECOUNT := 1 , AR := 0 , R := ROW-DUPSDONE;
570           ATOMLIST[R] := ATOMLIST[ROW] ;
571           RINGATOMTEST[R] := RINGATOMTEST[ROW] ;
572           UNITS[R] := UNITS[ROW] ;
573           'IF' ROW > R 'THEN' UNITS[ROW] := " " 'FI' ;
574           'REF'[]'INT' ADJROW = ADJACENCY[R, ] ;
575           'FOR' COLUMN 'TO' ATOMNUM 'DO'
576             'IF' COLUMN = E2[ECOUNT] 'THEN' ECOUNT 'PLUS' 1
577             'ELSF' CTROW[COLUMN] # 0 'THEN'
578               'INT' D = COLUMN - ECOUNT + 1 ;
579               'IF' AR # 4 'THEN' ADJROW[AR 'PLUS' 1] := D ;
580                 'IF' D < R 'THEN' BOND TABLE[BONDCOUNT 'PLUS' 1] := (D,R) 'FI'
581               'ELSE' FAILURE := "AR = 5" ; STRANGEATOM 'C' THE ICI PROGRAM
582                 SOMTIMES GIVES RING FUSION ATOMS AN ADDITIONAL SUBSTITUENT'C' 'FI'
583             'FI'
584         'ELSE' DUPSDONE 'PLUS' 1 'FI'
585         'END' ;
586
587       ATOMCOUNT := ATOMNUM-E ;
588       'FOR' X 'TO' ATOMCOUNT 'DO'
589         'BEGIN'
590           'C' SET UP MOLECULAR FORMULAE 'C'
591           'INT' J = 'CASE'(('ABS' ATOMLIST[X])-34) 'IN'
592                   1,0,2,3,4,0,5,0,0,0,0,6,7,8,0,0,9 'ESAC' ;
593           MOLFORM[J] 'PLUS' 1
594         'END' ;
```

```
595       'FOR' X 'TO' ATOMCOUNT 'DO'
596         'BEGIN'
597             'C' DETERMINE ATOMIC CONNECTIVITY VALUES 'C'
598           'REF''INT' CX = CONNECTIVITY[X] ; CX :=
599           ('INT' M := 0 ;
600           'REF'[]'INT' AX = ADJACENCY[X] ;
601           'FOR' Y 'TO' 4 'WHILE' AX[Y] # 0 'DO'
602             'IF' AX[Y] <= ATOMCOUNT 'THEN' M 'PLUS' 1
603             'ELSE' FAILURE := "NUMBERING FAILURE" ; STRANGEATOM 'FI' ; M )
604           ;'IF' CX = 0 'THEN' FAILURE := "ZERO VALENT ATOM" ; STRANGEATOM 'FI'
605         'END' ;
606
607   EXIT :
608       OK
609   'END' ;
610
611
612
613   'PROC' MATCH STRUCTURES = 'BOOL' :
614   'BEGIN'
615
616             'C' AUTHOR : PETER WILLETT , PGSLIS , 1977 .
617             THIS PROGRAM CARRIES OUT AN HEURISTIC STRUCTURE-MATCHING OF
618             TWO ADJACENCY MATRICES USING AN ADAPTION OF THE MORGAN ALGORITHM .
619             THE PROGRAM IDENTIFIES PAIRS OF ATOMS , ONE IN THE REACTANT AND
620             ONE IN THE PRODUCT , WHICH ARE THE MOST SIMILAR ONE-TO-ANOTHER .
621             THE RADIUS OF SIMILARITY , AS DETERMINED BY THE NUMBER OF
622             ITERATIONS OF THE MORGAN ALGORITHM FOR WHICH THE ATOMS HAVE
623             IDENTICAL PROPERTY VALUES , IS ASSUMED TO BE IDENTICAL AND ALL ATOMS
624             WITHIN THESE TWO AREAS MAY HENCE BE DELETED FROM FURTHUR CONSID-
625             ERATION . THE PROCEDURE ITERATES , MATCHING SMALLER AND SMALLER AREAS
626             , EITHER UNTIL AN AMBIGUITY IS DISCOVERED OR UNTIL THE MATCH RADIUS
627             FALLS TO BELOW 3 IE TWO BONDS DISTANT 'C'
628
629
```

```
630       'BOOL' OK := 'TRUE' ;
631       'LONG''INT' ZERO = 'LONG' 0 , TWO = 'LONG' 2 ;
632       'STRING' MATCHFAILURE ;
633       'PROC' MATCHFAILROUTINE = 'VOID' :
634         'BEGIN'
635           PUT(ANALYSISFAILURE,(NEWLINE,RULN,NEWLINE,PWLN,NEWLINE,MATCHFAILURE));
636           OK:= 'FALSE' ;
637           'GOTO' END MATCH STRUCTURES
638         'END' ;
639             'C' SET UP INITIAL PROPERTY VALUES 'C'
640       [1:RATOMCOUNT,1:RATOMCOUNT]'LONG''INT' RATOMPROP ; 'CLEAR' RATOMPROP ;
641       'REF'[]'LONG''INT' RAP = RATOMPROP[,1] ;
642       [1:PATOMCOUNT,1:PATOMCOUNT]'LONG''INT' PATOMPROP ; 'CLEAR' PATOMPROP ;
643       'REF'[]'LONG''INT' PAP = PATOMPROP[,1] ;
644       'FOR' X 'TO' RATOMCOUNT 'DO'
645         RAP[X] := 'LENG'(10*('ABS'RUNITS[X]-33) + RCONNECTIVITY[X] ) ;
646       'FOR' X 'TO' PATOMCOUNT 'DO'
647         PAP[X] := 'LENG'(10*('ABS'PUNITS[X]-33) + PCONNECTIVITY[X] ) ;
648
649       'PROC' EXTEND = ('INT' LEVEL)'VOID' :
650         'BEGIN'
651             'C' THIS IS BASICALLY THE MORGAN ALGORITHM EXCEPT THAT THE
652             INITIAL PROPERTY VALUE IS DERIVED FROM THE "UNITS" VALUE RATHER
653             THAN JUST THE CONNECTIVITY 'C'
654             'C' "REVIVE" AND "RESET REPORT" ARE INCLUDED IN CASE AN OVERFLOW
655             OCCURS DURING THE CALCULATION OF A PROPRTY VALUE 'C'
656           'REF'[]'LONG''INT'
657             RAPCURRENT = RATOMPROP[,LEVEL] ,RAP = RATOMPROP[,LEVEL-1] ,
658             PAPCURRENT = PATOMPROP[,LEVEL] , PAP = PATOMPROP[,LEVEL-1] ;
659           'LONG''INT' SUM ;
660           'PROC' OVERFLOWFAILURE = 'VOID' :
661             'BEGIN'
662               OVERFLOWINMATCH 'PLUS' 1 ;
663               MATCHFAILURE := "OVERFLOW SET" ;
664               MATCHFAILROUTINE
```

```
665             'END' ;
666           'FOR' X 'TO' RATOMCOUNT 'DO'
667             'BEGIN'
668               'REF'[]'INT' RADJX = RADJACENCY[X] ;
669               SUM := ZERO ;
670               REVIVE(OVERFLOWFAILURE) ;
671               RAPCURRENT[X] := TWO*RAP[X] + ('FOR' C 'TO' RCONNECTIVITY[X] 'DO'
672                                   SUM 'PLUS' RAP[RADJX[C]] ; SUM ) :
673               RESET REPORT
674             'END' ;
675           'FOR' X 'TO' PATOMCOUNT 'DO'
676             'BEGIN'
677               'REF'[]'INT' PADJX = PADJACENCY[X] ;
678               SUM := ZERO ;
679               REVIVE(OVERFLOWFAILURE) ;
680               PAPCURRENT[X] := TWO*PAP[X] + ('FOR' C 'TO' PCONNECTIVITY[X] 'DO'
681                                   SUM 'PLUS' PAP[PADJX[C]] ; SUM ) :
682               RESET REPORT
683           'END'
684         'END' ;
685
686       EXTEND(2) ; EXTEND(3) ;
687       'REF'[]'LONG''INT' FIRSTRAP = RATOMPROP[,3] , FIRSTPAP = PATOMPROP[,3] ;
688       'MODE''MATCH' = 'STRUCT'('INT' MAX,NUM,[1:0'FLEX']'INT' MATCHES) ;
689       'MATCH' DUPLICATES ;
690       [1:RATOMCOUNT]'MATCH' REACATOM ; 'INT' MAXIMAL := 0 :
691       [1:RATOMCOUNT,1:PATOMCOUNT]'INT' MATCHARRAY ; 'CLEAR' MATCHARRAY :
692       [1:PATOMCOUNT]'INT' SINGLERPMAP ; 'CLEAR' SINGLERPMAP :
693       'INT' RL := 0 , PL := 0 ;
694       [1:RATOMCOUNT]'INT' RLEFT ; [1:PATOMCOUNT]'INT' PLEFT :
695       [1:RATOMCOUNT,1:2]'INT' MATCHPAIR ; 'INT' MP := 0 :
696       'INT' MAXIMALATOMS ;
697       'BOOL' FOUNDMAPPING := 'FALSE' :
698
699       'PROC' MATCHATOMS = 'VOID' :
```

```
700       'BEGIN'
701             'C' "RLEFT" AND "PLEFT" ARE ARRAYS CONTAINING THE ATOMS NOT
702             PREVIOUSLY DELETED . THESE TWO SETS ARE MATCHED AGAINST EACH OTHER
703             USING THE THIRD-ORDER ATOM PROPERTIES AS THE INITIAL SET-PARTITIONING
704             CRITERION 'C'
705           'INT' J = RL , K = PL ; RL := PL := 0 ;
706           'FOR' X 'TO' K 'DO'
707             ('INT' M = PLEFT[X] ;
708             'IF' 'NOT' PDELETED[M] 'THEN' PLEFT[PL 'PLUS' 1] := M 'FI' ) ;
709           'FOR' X 'TO' J 'DO'
710             ('INT' R = RLEFT[X] ;
711             'IF''NOT' RDELETED[R] 'THEN'
712               'REF'[]'INT' MA = MATCHARRAY[R] ;
713               [1:PL]'INT' POSS ;
714               'INT' POSSCOUNT := 1 , M := 1 ;
715               MAX 'OF' REACATOM[R] := 0 ;
716               RLEFT[RL 'PLUS' 1] := R ;
717               'C' THIS LOOP MATCHES EACH PODUCT ATOM(P) AGAINST THE CURRENT
718               REACTANT ATOM(R) FOR AS MANY LEVELS AS POSSIBLE . THE VALUE
719               "LEVELCOUNT" IS THEN COMPARED WITH "M" , THE MAXIMUM VALUE FOR THAT
720               ATOM AND IF >= "M" THE APPROPRIATE INFORMATION STORED 'C'
721               'FOR' Y 'TO' PL 'DO'
722               'BEGIN'
723                 'INT' P = PLEFT[Y] ;
724                 'INT' LC = MA[P] ;
725                 'IF' LC > 2 'THEN'
726                   'IF' LC = M 'THEN' POSS[POSSCOUNT] := P ; POSSCOUNT 'PLUS' 1
727                   'ELSF' LC > M 'THEN' POSS[1] := P ; POSSCOUNT := 2 ;
728                     M := LC 'FI' ;
729                   REACATOM[R] := (M,POSSCOUNT-1,POSS[1:POSSCOUNT-1])
730                 'FI'
731               'END' ;
732               'IF' MAXIMAL < M 'THEN' MAXIMAL := M 'FI'
733             'FI' ) ;
734         'FOR' X 'TO' RL 'DO'
```

```
        ('MATCH' M = REACATOM[RLEFT[X]] ;
        'IF' NUM 'OF' M = 1 'AND' MAX 'OF' M = MAXIMAL 'THEN'
          SINGLERPMAP[(MATCHES 'OF' M)[1]] 'PLUS' 1 'FI' )
    'END' ;

    'PROC' DELETE = 'VOID' :
          'C' THIS PROCEDURE OPERATES UPON "MATCHARRAY" , EACH ELEMENT OF
          WHICH CONSISTS OF A REACTANT AND PRODUCT ATOM THAT HAVE BEEN
          JUDGED TO BE EQUIVALENT AT A MATCH RADIUS OF MAXIMAL - 1 BONDS .
          ALL ATOMS WITHIN THE 2 CIRCULAR SUBSTRUCTURES OF RADIUS MAXIMAL - 2
          ARE DELETED BY ITERATIVELY UPDATING "DELETED REACTANT" AND
          "DELETED PRODUCT" 'C'
      'BEGIN'
        [1:RATOMCOUNT]'BOOL' NBALLR , CURRENTR ;
        [1:PATOMCOUNT]'BOOL' NBALLP , CURRENTP ;
        'CLEAR' NBALLR ; 'CLEAR' NBALLP ;
        'REF'[]'INT' MP1 = MATCHPAIR[,1] , MP2 = MATCHPAIR[,2] ;
        MAXIMALATOMS 'MINUS' MP ;
        FOUNDMAPPING := 'TRUE' ;
        'FOR' X 'TO' MP 'DO'
          NBALLR[MP1[X]] := NBALLP[MP2[X]] := 'TRUE' ;
        'TO' MAXIMAL=2 'DO''BEGIN'
            'CLEAR' CURRENTR ; 'CLEAR' CURRENTP ;
            'FOR' X 'TO' RATOMCOUNT 'DO'
              'IF' NBALLR[X] 'THEN''REF'[]'INT' AX = RADJACENCY[X] ; 'FOR' Y
                'TO' RCONNECTIVITY[X] 'DO' CURRENTR[AX[Y]] := 'TRUE''FI' ;
            'FOR' X 'TO' PATOMCOUNT 'DO'
              'IF' NBALLP[X] 'THEN''REF'[]'INT' AX = PADJACENCY[X] ; 'FOR' Y
              'TO' PCONNECTIVITY[X] 'DO' CURRENTP[AX[Y]] := 'TRUE''FI' ;
            'FOR' X 'TO' RATOMCOUNT 'DO'
              NBALLR[X] := NBALLR[X] 'OR' CURRENTR[X] ;
            'FOR' X 'TO' PATOMCOUNT 'DO'
              NBALLP[X] := NBALLP[X] 'OR' CURRENTP[X]
          'END' ;
        'FOR' X 'TO' RATOMCOUNT 'DO'
```

```
            RDELETED[X] := RDELETED[X] 'OR' NBALLR[X] ;
          'FOR' X 'TO' PATOMCOUNT 'DO'
            PDELETED[X] := NBALLP[X] 'OR' PDELETED[X]
        'END' ;

      MAXIMAL := 0 ;
      ([1:RATOMCOUNT]'BOOL' RPOSS ; 'CLEAR' RPOSS ;
      [1:PATOMCOUNT]'BOOL' PPOSS ; 'CLEAR' PPOSS ;
      'INT' MAXPOSSIBLEMATCHRADIUS = ('IF' RATOMCOUNT > PATOMCOUNT 'THEN'
        PATOMCOUNT 'ELSE' RATOMCOUNT 'FI') - 4 ;
      'FOR' X 'TO' RATOMCOUNT 'DO'
        'BEGIN'
          'C' THIS LOOP DOES THE INITIAL ATOM-ATOM MATCHING . FIRST , SECOND
          AND THIRD ORDER ATOM PROPERTIES HAVE ALREADY BEEN CALCULATED ; A
          MATCH RADIUS < 3 IS NOT CONSIDERED SO THE INITIAL CRITERION FOR
          MATCH-SET GENERATION IS THAT THE THIRD-ORDER PROPERTY VALUES -
          ISOLATED IN THE ARRAYS "FIRSTRAP" AND "FIRSTPAP" - ARE IDENTICAL .
          IF THIS CONDITION HOLDS , THE PROPERTY VALUES ARE INCREMENTED - VIA
          THE PROCEDURE "EXTEND" - FOR AS LONG AS POSSIBLE . MAXIMAL IS THE
          CURRENT MAXIMAL MATCH RADIUS FOR ALL ATOMS , M THAT FOR THE
          CURRENT REACTANT ATOM 'C'
        'REF'[]'LONG''INT' RAR = RATOMPROP[X] ;
        [1:PATOMCOUNT]'INT' POSS ; 'CLEAR' POSS ;
        'REF'[]'INT' MA = MATCHARRAY[X] ;
        'INT' POSSCOUNT := 1 , M := 1 ;
        MAX 'OF' REACATOM[X] := 0 ;
        'FOR' Y 'TO' PATOMCOUNT 'DO'
          'BEGIN'
            'REF'[]'LONG''INT' PAP = PATOMPROP[Y] ;
            'IF' ('BOOL' B := 'TRUE' ; 'FOR' X 'TO' 3 'WHILE' B 'DO'
              B := RAR[X] = PAP[X] ; B) 'THEN' RPOSS[X] := PPOSS[Y] := 'TRUE' ;
              'BEGIN'
                'REF'[]'LONG''INT' PAP = PATOMPROP[Y] ;
                'INT' LC := 3 , LEVELCOUNT := 4 ;
                'TO' MAXPOSSIBLEMATCHRADIUS 'WHILE' RAR[LC] = PAP[LC] 'DO'
```

```
805                    ('IF' RAR[LEVELCOUNT] = ZERO 'THEN' EXTEND(LEVELCOUNT)'FI' :
806                    LEVELCOUNT := (LC 'PLUS' 1) + 1 ) ;
807                  MA[Y] := LC 'MINUS' 1 ;
808                  'IF' LC = M 'THEN'
809                    POSS[POSSCOUNT] := Y ; POSSCOUNT 'PLUS' 1
810                  'ELSF' LC > M 'THEN' POSS[1] := Y ; POSSCOUNT := 2 ;
811                    M := LC 'FI'
812                'END' ;
813                REACATOM[X] := (M,POSSCOUNT=1,POSS[1:POSSCOUNT=1])
814              'FI'
815            'END' ;
816          'IF' MAXIMAL < M 'THEN' MAXIMAL := M 'FI'
817        'END' ;
818      'FOR' X 'TO' RATOMCOUNT 'DO'
819        'IF' RPOSS[X] 'THEN''MATCH' M = REACATOM[X] : RLEFT[RL 'PLUS' 1] := X ;
820          'IF' NUM 'OF' M = 1 'AND' MAX 'OF' M = MAXIMAL 'THEN'
821            SINGLERPMAP[(MATCHES 'OF' M)[1]] 'PLUS' 1 'FI'
822        'FI' ;
823      'FOR' Y 'TO' PATOMCOUNT 'DO'
824        'IF' PPOSS[Y] 'THEN' PLEFT[PL 'PLUS' 1] := Y 'FI' ) ;
825
826      'IF' MAXIMAL >= 3 'THEN''BOOL' ELIMINATE := 'TRUE' : 'WHILE' ELIMINATE 'DO'
827        'BEGIN'
828          'BOOL' CHECK := 'FALSE' ;
829          'INT' MA := 0 ;
830          MAXIMALATOMS := 0 ;
831          DUPLICATES := (MAXIMAL,0,RLEFT) ;
832          'FOR' X 'TO' RL 'DO''BEGIN'
833              'C' SEARCH THROUGH THE REMAINING REACTANT ATOMS AND NOTE
834              THOSE HAVING A MATCH AT THE MAXIMAL MATCH RADIUS 'C'
835              'INT' RLX = RLEFT[X] ;
836              'REF''MATCH' RARLX = REACATOM[RLX] ;
837              'IF' MAX 'OF' RARLX = MAXIMAL 'THEN' MAXIMALATOMS 'PLUS' 1 ;
838                'IF' NUM 'OF' RARLX = 1 'THEN'
839                  'INT' Z = (MATCHES 'OF' RARLX)[1] :
```

```
840                 'IF' SINGLERPMAP[Z] = 1 'THEN'
841                   MATCHPAIR[MP 'PLUS' 1] := (RLX,Z) ; CHECK := 'TRUE''FI'
842               'ELSE' (MATCHES 'OF' DUPLICATES)[NUM 'OF' DUPLICATES 'PLUS' 1]
843                 := RLX 'FI'
844             'FI'
845           'END' ;
846         'IF' MP > 0 'THEN' DELETE ; MP := 0 'FI' ;
847           'C' ANY MAXIMAL REACTANT ATOMS HAVING MORE THAN ONE POSSIBLE
848           MAPPING(THESE REACTANT ATOMS ARE STORED IN "MATCHES OF DUPLICATES" )
849           ARE CHECKED TO SEE WHETHER ANY OF THE PRODUCT ATOMS HAVE BEEN DELETED
850           IN THE FIRST SET OF MATCHINGS AT THE CURRENT MATCH RADIUS 'C'
851         'IF' MAXIMALATOMS > 0 'AND' NUM 'OF' DUPLICATES > 0 'THEN'
852           'REF''INT' M = NUM 'OF' DUPLICATES ; 'REF'[]'INT' MOD =
853           (MATCHES 'OF' DUPLICATES)[1:M] ;
854           'IF' CHECK 'THEN''FOR' X 'TO' M 'DO''BEGIN'
855               'INT' M := (MATCHES 'OF' DUPLICATES)[X] ;
856               'REF''MATCH' RAM = REACATOM[(MATCHES 'OF' DUPLICATES)[X]] ;
857               'REF''INT' MODX = (MATCHES 'OF' DUPLICATES)[X] ;
858               'IF' RDELETED[M] 'THEN'
859                 MODX := 999
860               'ELSE''INT' P := 0 ;
861                 'FOR' Y 'TO' NUM 'OF' RAM 'DO'
862                   'IF''NOT' PDELETED[(MATCHES 'OF' RAM)[Y]] 'THEN'
863                     (MATCHES 'OF' RAM)[P 'PLUS' 1] := (MATCHES 'OF' RAM)[Y]'FI';
864                 NUM 'OF' RAM := P ;
865                 'IF' P = 1 'THEN' MATCHPAIR[MP 'PLUS' 1] := (M,(MATCHES
866                   'OF' RAM)[1]) ;
867                   CHECK := 'TRUE' ; MODX := 999 'FI'
868               'FI'
869             'END' ; 'IF' MP > 0 'THEN' DELETE ; MP := 0 'FI'
870           'FI' ;
871           'FOR' X 'TO' M 'DO'
872             'IF' MOD[X] # 999 'THEN'
873               'REF''INT' MODX = MOD[X] ;
874               'REF''MATCH' RAMODX = REACATOM[MODX] ;
```

```
875                 'IF' NUM 'OF' RAMODX > 0 'THEN'
876                   'INT' J := 0 , NOR := NUM 'OF' RAMODX ;
877                   'REF'[]'INT' MORX = MATCHES 'OF' RAMODX :
878                   [1:NOR*NOR]'INT' ANALOGUES ;
879                   'C' THE NEXT LOOP CHECKS ALL REMAINING MAXIMAL REACTANT ATOMS
880                   -"MODY"- HAVING MORE THAN ONE POSSIBLE MAPPING TO SEE WHETHER
881                   THEY POSSESS THE SAME MATCH-SET AS THE CURRENT ONE -"MODX"- ;
882                   IF SO , "MODY" IS STORED IN THE LIST OF "ANALOGUES" 'C'
883                   'FOR' Y 'TO' M 'DO'
884                     'BEGIN'
885                       'REF''INT' MODY = MOD[Y] ;
886                       'IF' MODY # 999 'AND' MODY # MODX 'THEN'
887                         'REF''MATCH' RAMODY = REACATOM[MODY] ;
888                         'IF' (NUM 'OF' RAMODX = NUM 'OF' RAMODY) 'AND'
889                           ((MATCHES 'OF' RAMODX)[1:NOR] 'EQUALS'
890                            (MATCHES 'OF' RAMODY)[1:NOR]) 'THEN'
891                           ANALOGUES[J 'PLUS' 1] := MODY 'FI'
892                       'FI'
893                     'END' ;
894                   'C' IF THE NUMBER OF ANALOGUES IS EQUAL TO THE SIZE OF THE
895                   MATCH-SET THEN THE REACTANT ATOM , ITS ANALOGUES AND THE
896                   MATCH-SET ARE ALL ELIMINATED 'C'
897                   'IF' J = NOR - 1 'THEN'
898                     'FOR' Z 'TO' J 'DO''BEGIN'
899                         'REF''INT' AZ = ANALOGUES[Z] ;
900                         MATCHPAIR[MP 'PLUS' 1] := (AZ,MORX[Z]) ; CHECK := 'TRUE' ;
901                         'FOR' X 'TO' M 'DO'
902                           'IF' MOD[X] = AZ 'THEN' MOD[X] := 999 'FI'
903                       'END' ;
904                     MATCHPAIR[MP 'PLUS' 1] := (MODX,MORX[J+1]) ; CHECK := 'TRUE' ;
905                     'FOR' X 'TO' M 'DO'
906                       'IF' MOD[X] = MODX 'THEN' MOD[X] := 999 'FI'
907                   'FI'
908                 'FI'
909               'FI' ;
```

```
910           'IF' MP > 0 'THEN' DELETE ; MP := 0 'FI'
911         'FI' ;
912           'C' THE ALGORITHM TERMINATES EITHER IF THE MAXIMAL MATCH RADIUS < 3
913           OR IF THERE ARE STILL REMAINING ATOMS WITH THE CURRENT MAXIMAL
914           MATCH RADIUS 'C'
915         ELIMINATE :=
916           'IF' MAXIMALATOMS # 0 'OR' MAXIMAL = 3 'THEN''FALSE'
917           'ELSE' MAXIMAL := 0 ; MATCHATOMS ;
918             'IF' MAXIMAL < 3 'THEN''FALSE''ELSE' CHECK 'FI'
919           'FI'
920       'END'
921     'FI' ;
922
923           'C' THIS ROUTINE DETECTS SOLITARY REACTANT ATOMS IN THE REACTION
924           SITES IE THOSE WHICH ARE NOT ATTACHED TO ANY OTHER SITE ATOMS ; IF ONE
925           IS FOUND , AN ATTEMPT IS MADE TO MATCH IT WITH AN ANALOGOUS PRODUCT
926           ATOM 'C'
927     'IF' FOUNDMAPPING 'THEN''FOR' X 'TO' RATOMCOUNT 'DO'
928       'IF''NOT' RDELETED[X] 'THEN'
929         'REF'[]'INT' RAX = RADJACENCY[X] ;
930         'BOOL' B := 'TRUE' ;
931         'FOR' Y 'TO' RCONNECTIVITY[X] 'WHILE' B 'DO'
932           B := RDELETED[RAX[Y]] ;
933         'IF' B 'THEN'
934           'BOOL' NOMATCH := 'TRUE' ;
935           'CHAR' R = RUNITS[X] ;
936           'FOR' A 'TO' PATOMCOUNT 'WHILE' NOMATCH 'DO'
937             'IF''NOT' PDELETED[A] 'AND' PUNITS[A] = R 'THEN'
938               'REF'[]'INT' PAA = PADJACENCY[A] ;
939               'BOOL' B := 'TRUE' ;
940               'FOR' C 'TO' PCONNECTIVITY[A] 'WHILE' B 'DO'
941                 B := PDELETED[PAA[C]] ;
942               'IF' B 'THEN'
943                 RDELETED[X] := PDELETED[A] := 'TRUE' ;
944                 NOMATCH := 'FALSE'
```

```
945                 'FI'
946               'FI'
947             'FI'
948           'FI'
949         'FI' ;
950
951             'C' CHECK THAT AT LEAST ONE MATCH HAS BEEN OBTAINED AND THAT NOT
952             ALL THE ATOMS ON EITHER SIDE OF THE EQUATION HAVE BEEN ELIMINATED 'C'
953         'IF''NOT' FOUNDMAPPING 'THEN' OK := 'FALSE' ; NOMATCH 'PLUS' 1 ;
954           MATCHFAILURE := "NO ATOMS MATCHED" ; MATCHFAILROUTINE
955         'ELSE''INT' LEFT := 0 ;
956           'PROC' SETATOMBITS = ('INT' B) :
957             'BEGIN'
958               'IF' LEFT > 1 'THEN''FOR' X 'FROM' B 'TO''IF' LEFT>8 'THEN' (LEFT+7)
959                 'ELSE' (B+LEFT-2)'FI''DO'
960                 MINIMUM := X 'SET' MINIMUM
961               'ELSE' MATCHFAILURE := 'IF' REACTANT 'THEN'
962                 "LESS THAN TWO ANALYSIS REACTANT ATOMS" 'ELSE'
963                 "LESS THAN TWO ANALYSIS PRODUCT ATOMS" 'FI'; ALLMATCH 'PLUS' 1 ;
964                 MATCHFAILROUTINE
965               'FI'
966             'END' ;
967           'FOR' X 'TO' RATOMCOUNT 'DO'
968             'IF''NOT' RDELETED[X] 'THEN' LEFT 'PLUS' 1 'FI' ;
969           SETATOMBITS(9) ; RPERCENT[(LEFT*100)'/'RATOMCOUNT]'PLUS' 1 :
970           LEFT := 0 ; 'FOR' X 'TO' PATOMCOUNT 'DO'
971             'IF''NOT' PDELETED[X] 'THEN' LEFT 'PLUS' 1 'FI' ;
972           SETATOMBITS(17) ; PPERCENT[(LEFT*100)'/'PATOMCOUNT]'PLUS' 1
973       'FI' ;
974
975   END MATCH STRUCTURES :
976       OK
977   'END' ;
978
979
```

```
'PROC' RING ANALYSIS SCREENS = ('REF'[]'BITS' ARSBIT) :
  'BEGIN'
    'C' THIS PROCEDURE READS THE ANALYSIS FRAGMENTS OUTPUT FROM
    THE WLN PROGRAM.  THE FRAGMENTS, UPTO 10 IN NUMBER, ARE 20 CHAR. LONG
    AND THE RINGS START WITH #,L,T OR R.  THREE LEVELS OF DESCRIPTION
    ARE USED :
    (1) SIZE
    (2) NUMBER OF HETEROATOMS.
    (3) A NUMBER DESCRIBING THE HETEROATOM TYPES OR, IF NONE ARE PRESENT,
    THE SATURATION.  IN EITHER CASE PLUS ONE MILLION IF THE RING IS
    FUSED.  BITS ARE ALSO SET IN MINIMUM 'C'
    'INT' P := PA := 'IF' REACTANT 'THEN' 1801 'ELSE' 2001 'FI' ;
    'INT' RINGCOUNT := 0 ;
    'PROC' SETRINGBIT = ('INT' B) :
      'BEGIN'
        'IF' RINGCOUNT > 1 'THEN'
          MINIMUM := (B+1)'SET' MINIMUM ; MINIMUM := B 'SET' MINIMUM
        'ELSF' RINGCOUNT = 1 'THEN' MINIMUM := B 'SET' MINIMUM 'FI'
      'END' ;
    'TO' 10 'WHILE' C[P] # " " 'DO'
      'BEGIN'
        'REF''CHAR' CHAR = C[P] ;
        'IF' CHAR = "#" 'OR' CHAR = "T" 'OR' CHAR = "L" 'THEN'
          [1:3]'INT' RING ;
          []'CHAR' HET = ("V","O","S","N","M","K","P") ;
          [1:7]'INT' FREQ ; 'CLEAR' FREQ ;
          'INT' PLIMIT = P + 16 ;
          'WHILE' C[P] > "9" 'DO' P 'PLUS' 1 ;
          RING[1] := 'ABS' C[P] ;
          P 'PLUS' 1 ;
          'WHILE' ( P < PLIMIT 'AND' C[P] = " ")'DO''BEGIN'
              'BOOL' B := 'TRUE' ;
              'REF''CHAR' CHAR = C[P 'PLUS' 2] ;
              'FOR' X 'TO' 7 'WHILE' B 'DO'
                'IF' CHAR = HET[X] 'THEN' B := 'FALSE' ;
```

```
1015                    FREQ[X] 'PLUS' 1 'FI' ;
1016                  P 'PLUS' 1
1017                'END' ;
1018               FREQ[4] 'PLUS' FREQ[5] 'PLUS' FREQ[6] ; FREQ[5] := FREQ[7] :
1019               'INT' I := 0 , J := 0 ;
1020               'FOR' X 'TO' 5 'DO' (I 'PLUS' FREQ[X] ; J 'PLUS'
1021                 (10↑(X-1))*FREQ[X]) ;
1022               RING[2] := 'IF' I # 0 'THEN' I 'ELSE' 99 'FI' ;
1023               'IF' I # 0 'THEN' RING[3] := J
1024               'ELSE''WHILE' C[P] # "J" 'DO' P 'PLUS' 1 : P 'MINUS' 1 :
1025                 RING[3] := 'IF' C[P] = "T" 'THEN' 1 'ELSE' 2 'FI'
1026               'FI' ;
1027               'IF' CHAR = "#" 'THEN' RING[3] 'PLUS' 1000000 'FI' :
1028               SEARCH(RING,ANALRINGINDEX,ANALRINGPARENT,ANALRINGSCREEN,
1029                 ARSBIT) ;
1030               RINGCOUNT 'PLUS' 1 ;
1031               P := PA 'PLUS' 20
1032             'ELSE' CHAR = "R" 'THEN'
1033               ARSBIT[1] := 23 'SET' ARSBIT[1] ; RINGCOUNT 'PLUS' 1
1034             'ELSE' P := PA 'PLUS' 20 'FI'
1035           'END'
1036         ; 'IF' REACTANT 'THEN' SETRINGBIT(3) 'ELSE' SETRINGBIT(7) 'FI'
1037       'END' ;
1038
1039     'PROC' ATOMBONDSCREENS = ('REF'[]'CHAR' UNITS,ATOMLIST,'INT' ATOMCOUNT,
1040       BONDCOUNT,'REF'[]'BOOL' RINGATOMTEST,DELETED,'REF'[,]'INT' ADJACENCY,
1041       BONDTABLE,'REF'[]'INT' CONNECTIVITY,'REF'[]'BITS' ASBIT,AASBIT,BSBIT,
1042       ABSBIT) :
1043       'BEGIN'
1044         [1:6]'INT' BONDPROP ;
1045         [1:ATOMCOUNT,1:7]'INT' ATOMPROP ;
1046         'REF'[]'INT' AP1 = ATOMPROP[,1] , AP2 = ATOMPROP[,2] ,
1047           AP3 = ATOMPROP[,3] ;
1048         'REF'[,]'INT' AP37 = ATOMPROP[,3:7] ;
1049         'FOR' X 'TO' ATOMCOUNT 'DO''BEGIN'
```

```
1050            'C' THE FIRST THREE ATOMIC PROPERTY VALUES CORRESPOND TO THE
1051            ATOM TYPE, THE COORDINATED ATOM AND THE BONDED ATOM. SUBSEQUENT
1052            VALUES ARE CALCULATED USING THE MORGAN ALGORITHM 'C'
1053            AP1[X] := 'CASE''ABS' ATOMLIST[X]-32 'IN'
1054                      0,0,23,0,5,7,11,0,3,0,0,0,0,19,17,0,0,0,13 'ESAC' ;
1055            AP2[X] := AP1[X] + CONNECTIVITY[X] ; AP3[X] := 'ABS' UNITS[X]
1056          'END' ;
1057        'FOR' X 'FROM' 4 'TO' 7 'DO'
1058          'BEGIN'
1059            'REF'[]'INT' AP = ATOMPROP[,X] , LASTAP = ATOMPROP[,X-3] ;
1060            'FOR' Y 'TO' ATOMCOUNT 'DO'
1061              AP[Y] := 3*LASTAP[Y] + ('INT' SUM := 0 ; 'REF'[]'INT' A =
1062                                      ADJACENCY[Y] ;
1063                                      'FOR' C 'TO' CONNECTIVITY[Y] 'DO'
1064                                        SUM 'PLUS' LASTAP[A[C]] ;
1065                                      SUM )
1066          'END' ;
1067        'FOR' X 'TO' ATOMCOUNT 'DO''BEGIN'
1068            'REF'[]'INT' APX = AP3[X] ;
1069            'IF' RINGATOMTEST[X] 'THEN' AP1[X] 'PLUS' 100; AP2[X] 'PLUS' 100 ;
1070              AP3[X] 'PLUS' 100 'FI' ;
1071            'C' SET BITS IN THE MOLECULAR SCREENS AND, IF THE ATOM IS IN
1072            THE REACTION SITE, IN THE ANALYSIS SCREENS AS WELL 'C'
1073            SEARCH(APX,ATOMINDEX,ATOMPARENT,ATOMSCREEN,ASBIT) :
1074            'IF''NOT' DELETED[X] 'THEN' SEARCH(APX,ANALATOMINDEX,
1075              ANALATOMPARENT,ANALATOMSCREEN,AASBIT) 'FI'
1076          'END' ;
1077        'FOR' X 'TO' BONDCOUNT 'DO''BEGIN'
1078            'REF'[]'INT' BT = BONDTABLE[X] ;
1079            'REF'[]'INT' AP1 = ATOMPROP[BT[1]] , AP2 = ATOMPROP[BT[2]] ;
1080            'FOR' Y 'TO' 6 'DO' BONDPROP[Y] := AP1[Y]*AP2[Y] ;
1081            SEARCH(BONDPROP,BONDINDEX,BONDPARENT,BONDSCREEN,BSBIT) ;
1082            'C' THE ANALYSIS BOND SCREENS ARE SET ONLY IF BOTH THE ATOMS
1083              COMPRISING THE BOND ARE IN THE REACTION SITE 'C'
1084            'IF''NOT' DELETED[BT[1]] 'AND''NOT' DELETED[BT[2]] 'THEN'
```

```
1085              SEARCH(BONDPROP,ANALBONDINDEX,ANALBONDPARENT,ANALBONDSCREEN,
1086              ABSBIT) 'FI'
1087          'END'
1088      'END' ;
1089
1090  'PROC' MOLFORMSCREEN = ('INT' I,'REF'[]'INT' MOLFORM) :
1091    'BEGIN'
1092        'C' SET UP MOLECULAR FORMULA SCREENS . THE 23 BITS ARE USED AS BELOW :
1093        1-6 NUMBERS OF CARBONS (0-4,5-10,11-15..>25)
1094        7-11 INDIVIDUAL HALOGENS AND PHOSPHORUS
1095        12 GENERAL HALOGEN
1096        13-15 GENERAL HETEROATOMS (1,2,>2)
1097        16-18 NITROGEN ATOMS (1,2,>2)
1098        19-21 OXYGEN ATOMS (1,2,>2)
1099        22-23 SULPHUR ATOMS (1,>1)
1100        'C'
1101      'REF''BITS' MFS = (MOLFORMBITS 'OF' OUTPUTBUFFER)[1] ;
1102      'INT' HALOGEN := 0 , HETEROATOM := 0 ;
1103      'BOOL' B ;
1104      'INT' J := 'IF' I = 2 'OR' I = 4 'THEN' MOLFORM[1]
1105        'ELSE' (MOLFORM[1]'/'5)+1 'FI' ;
1106      'IF' J # 0 'THEN''FOR' X 'TO''IF' J < 6 'THEN' J 'ELSE' 6 'FI''DO'
1107        MFS := X 'SET' MFS 'FI' ;
1108      'FOR' X 'FROM' 2 'TO' 5 'DO'
1109        'IF' MOLFORM[X] > 0 'THEN' MFS := (X+5)'SET'MFS : HALOGEN 'PLUS'
1110        MOLFORM[X] 'FI' ;
1111      'IF' HALOGEN > 0 'THEN' MFS := 12'SET'MFS 'FI' ;
1112      'FOR' X 'FROM' 2 'TO' 9 'DO' HETEROATOM 'PLUS' MOLFORM[X] ;
1113      'IF' MOLFORM[8] > 0 'THEN' MFS := 11'SET'MFS 'FI' ;
1114      'IF' MOLFORM[6] > 0 'THEN''C' NITROGEN ATOM 'C'
1115        'FOR' X 'FROM' 16 'TO''CASE' MOLFORM[6] 'IN' 16,17 'OUT' 18 'ESAC''DO'
1116        MFS := X 'SET' MFS 'FI' ;
1117      'IF' MOLFORM[7] > 0 'THEN''FOR' X 'FROM' 19 'TO''CASE' MOLFORM[7]
1118        'IN' 19,20 'OUT' 21 'ESAC' 'DO' MFS := X 'SET' MFS 'FI' :
1119      J := MOLFORM[9] ; 'C' SULPHUR 'C'
```

```
1120          'IF' J = 1 'THEN' MFS := 22'SET'MFS ; HETEROATOM 'PLUS' 1
1121          'ELSF' J > 1 'THEN' MFS := 22 'SET' MFS ; MFS := 23 'SET' MFS 'FI' ;
1122          'IF' HETEROATOM > 0 'THEN' 'C' SET GENERAL HETEROATOM BITS 'C'
1123            'FOR' X 'FROM' 13 'TO''CASE' HETEROATOM 'IN' 13,14 'OUT' 15 'ESAC'
1124            'DO' MFS := X 'SET' MFS 'FI'
1125        'END' ;
1126
1127
1128
1129  'DO''BEGIN'
1130      'CLEAR' RINGBITS 'OF' OUTPUTBUFFER ; 'CLEAR' MOLATOMBITS 'OF'
1131        OUTPUTBUFFER ; 'CLEAR' ANALATOMBITS 'OF' OUTPUTBUFFER ; 'CLEAR'
1132        MOLBONDBITS 'OF' OUTPUTBUFFER ; 'CLEAR' ANALBONDBITS 'OF' OUTPUTBUFFER ;
1133      MINIMUM := 'BIN' 0 ;
1134      'CLEAR' MOLFORMBITS 'OF' OUTPUTBUFFER ;
1135      'CLEAR' RCONNECTIVITY ; 'CLEAR' PCONNECTIVITY ;
1136      'CLEAR' RADJACENCY ; 'CLEAR' PADJACENCY ;
1137      'CLEAR' RDELETED ; 'CLEAR' PDELETED ;
1138      'CLEAR' RUNITS ; 'CLEAR' PUNITS ;
1139      'CLEAR' RRINGATOMTEST ; 'CLEAR' PRINGATOMTEST ;
1140      'CLEAR' RBONDTABLE ; 'CLEAR' PBONDTABLE ;
1141      'CLEAR' RMOLFORM ; 'CLEAR' PMOLFORM ;
1142      'CLEAR' (RLN 'OF' OUTPUTBUFFER) ;
1143      REACTIONS 'PLUS' 1 ;
1144      MTRDB(1,(LENGTH 'OF' BUFFER)[1],576) ;
1145      'IF' C[439] = " " 'AND' C[1239] = " " 'THEN'
1146        'BOOL' OK := 'TRUE' ;
1147        PA := 1 ; REACTANT := 'TRUE' ;
1148        REACTANT := 'TRUE' ;
1149        OK := CREATE ADJACENCY TABLE(RADJACENCY,RBONDTABLE,RCONNECTIVITY,
1150          RMOLFORM,RATOMCOUNT,RBONDCOUNT,RRINGATOMTEST,RLN,RUNITS,RATOMLIST) ;
1151        REACTANT := 'FALSE' ; PA := 801 ;
1152        'IF' OK 'THEN' OK := CREATE ADJACENCY TABLE(PADJACENCY,PBONDTABLE,
1153          PCONNECTIVITY,PMOLFORM,PATOMCOUNT,PBONDCOUNT,PRINGATOMTEST,
1154          PLN,PUNITS,PATOMLIST) ;
```

```
1155          'IF' OK 'THEN' GENERATED ADJACENCY TABLE 'PLUS' 1 ;
1156            OK := MATCH STRUCTURES ;
1157            'IF' OK 'THEN'
1158              REACTANT := 'TRUE' ;
1159              RING ANALYSIS SCREENS((RINGBITS 'OF' OUTPUTBUFFER)[2]) ;
1160              REACTANT := 'FALSE' ;
1161              RING ANALYSIS SCREENS((RINGBITS 'OF' OUTPUTBUFFER)[4]) ;
1162              ATOMBONDSCREENS(RUNITS,RATOMLIST,RATOMCOUNT,RBONDCOUNT,
1163                RRINGATOMTEST,RDELETED,RADJACENCY,RBONDTABLE,RCONNECTIVITY,
1164                (MOLATOMBITS 'OF' OUTPUTBUFFER)[1],(ANALATOMBITS 'OF'
1165                OUTPUTBUFFER)[1],(MOLBONDBITS 'OF' OUTPUTBUFFER)[1],
1166                (ANALBONDBITS 'OF' OUTPUTBUFFER)[1]) ;
1167              ATOMBONDSCREENS(PUNITS,PATOMLIST,PATOMCOUNT,PBONDCOUNT,
1168                PRINGATOMTEST,PDELETED,PADJACENCY,PBONDTABLE,PCONNECTIVITY,
1169                (MOLATOMBITS 'OF' OUTPUTBUFFER)[2],(ANALATOMBITS 'OF'
1170                OUTPUTBUFFER)[2],(MOLBONDBITS 'OF' OUTPUTBUFFER)[2],
1171                (ANALBONDBITS 'OF' OUTPUTBUFFER)[2]) ;
1172              MOLFORMSCREEN(1,RMOLFORM) ; MOLFORMSCREEN(3,PMOLFORM) ;
1173              'FOR' X 'TO' 9 'DO'
1174                'BEGIN'
1175                  'REF''INT' RMX = RMOLFORM[X] , PMX = PMOLFORM[X] :
1176                  'IF' RMX > PMX 'THEN' RMX 'MINUS' PMX : PMX := 0
1177                  'ELSE' PMX 'MINUS' RMX ; RMX := 0 'FI'
1178                'END' ;
1179              MOLFORMSCREEN(2,RMOLFORM) ; MOLFORMSCREEN(4,PMOLFORM) :
1180              SUCCESS 'PLUS' 1 ;
1181              WLNFRAGMENTS 'OF' OUTPUTBUFFER := C[1801:2200] ;
1182              BIBDETAILS 'OF' OUTPUTBUFFER := C[2201:2220] ;
1183              MTWR(2,(LENGTH 'OF' OUTPUTBUFFER)[1]) :
1184              'SKIP'
1185            'FI'
1186          'FI'
1187        'FI'
1188      'FI'
1189  'END' ;
```

```
1190
1191  FINIS :
1192        PUT(MATRIXOUT,("NUMBER OF REACTIONS INPUT = ",REACTIONS,NEWLINE,NEWLINE,
1193          "NUMBER OF REACTIONS SUBMITTED TO THE STRUCTURE MATCHING ALGORITHM = ",
1194          GENERATED ADJACENCY TABLE",NEWLINE,NEWLINE,
1195          "NUMBER OF REACTIONS FOR WHICH NO ATOM/ATOM EQUIVALENCES WERE FOUND = ",
1196          NOMATCH,NEWLINE,NEWLINE,
1197          "NUMBER OF REACTIONS IN WHICH ALL THE ATOMS WERE ELIMINATED = ",
1198          ALLMATCH,NEWLINE,NEWLINE,"NUMBER OF REACTIONS FOR WHICH OVERFLOW SET = "
1199          ,OVERFLOWINMATCH,NEWLINE,NEWLINE)) ;
1200        PRINT((NEWLINE,REACTIONS,SUCCESS,NOMATCH,ALLMATCH)) :
1201        MTEND(1,"CLOSE") ; MTEND(2,"CLOSE") ;
1202        'SKIP'
1203  'END'
1204  'FINISH'
1205  ****
```

APPENDIX III.

A qualitative comparison of Wiswesser Line Notation descriptors of reactions and the Derwent Chemical Reaction Documentation Service.

David Bawden*, Trevor K. Devon, Frank T. Jackson and Sandra I. Wood,
(Pfizer Central Research, Sandwich, Kent)

and

Michael F. Lynch and Peter Willett,
(Postgraduate School of Librarianship and Information Science,
University of Sheffield, Western Bank, Sheffield, S10 2TN).

Summary.

Two methods of retrieving chemical reaction information are compared. One involves the generation of reaction descriptors automatically by an analysis of the Wiswesser Line Notation of the reacting molecules. The other, Derwent's Chemical Reaction Documentation Service (CRDS), involves manual indexing and uses a bond-change code to describe the reaction, with Ringcode for structural description. A series of reaction queries was searched using both systems: the results were qualitative and indicative of the general nature of the descriptions provided.

Both systems are found to perform effectively with queries involving a definite reaction site change. The WLN system gives greater precision in some cases, due to the varying levels of structural reqpresentation provided. CRDS is valuable where particular bond changes are specified, and could be valuable in synthetic planning. Neither system performs well with queries where no definite reaction site is specified, and both would require additional concept indexing for full effectiveness. The WLN system has a useful potential for producing printed indexes of reactions.

*To whom correspondence should be addressed.

Introduction.

The provision of access to chemical reaction information has been a continuing problem for chemical information workers, and a variety of approaches has been adopted.[1,2] One method involves the automatic generation of reaction descriptions from machine-readable representations of chemical structures. Such descriptions may then be searched by computer or used for the production of printed indexes. This approach is likely to be of particular value within computerised chemical information systems. Investigations along these lines have been carried out for some years at Sheffield, using both connection table and Wiswesser Line Notation (WLN) representations of structure.[3] This work has resulted in the development of a method of reaction analysis based on WLN.[4] The WLNs for the reactant and product molecules are fragmented algorithmically, the fragments compared and duplicates eliminated, and the remaining fragments then recombined to give a description of the reaction site. The fragments constituting the reaction site are the main entry points to the reaction file: further information may be obtained by considering the fuller reaction site notations and then the original WLNs. In a printed index these latter stages are carried out by scanning the entries under the appropriate reaction site fragment(s). In a computerised form, a string search procedure would be used on the reaction site notation and/or full WLN. At present this approach to reaction indexing is at an experimental stage.

A reaction documentation system based on structural concepts is, however, commercially available at the present time. This is the Chemical Reactions Documentation Service (CRDS) based on a system originally devised by the Pharma Dokumentation Ring.[6] This system describes reactions according to a representation of bonds formed and broken, derived from the coding used in Theilheimer's "Synthetic Methods" series.[7] Structures of reactant and product molecules are represented by the fragmentation code developed by the Pharma Dokumentations Ring (Ringcode)[8]. The service is amenable to computer searching in batch mode, the reactions being searched by the bond change codes and Ringcodes for the reactant and product structures. It has

been proposed by Derwent that this system will be made available on-line, with added keyword indexing.

A comparison of these two systems appears to be worthwhile in order to determine whether one type of reaction description is markedly superior to the other.

2. Methodology

There are major differences in the current state of implementation of the two systems. CRDS is a fully operational computerised system with facilities for searching on reaction conditions etc., and allowing searching of reactant and product structures using Ringcode; the Sheffield WLN system is still at an experimental stage and has a printed index output with manual scanning; thus, the provision for whole structure searching in the two systems is so different that, for example, relative precision figures would be meaningles For these reasons, and because the main objective of the study was a comparison of the basic reaction descriptions provided, rather than of overall system effectiveness, no formal quantitative evaluation was attempted. Rather, the aim was to produce a qualitative understanding of the strengths, weaknesses and potentialities of each method: quantitative evaluation would only be appropriate in the context of two fully operational systems. Ease of use and other human factors were not specifically examined, but had to be taken into account to some extent. Using printed tools, especially with a relatively small data-base, it is easy to scan a large proportion of the possible results. Some subjective judgement as to what would be realistic in a practical applicatio was therefore necessary.

The data base used for the evaluation consisted of 273 abstracts, chosen randomly from each of volumes 22, 24 and 30 of Theilheimer's "Synthetic Methods"; this series forms the bulk of the CRDS data base. Each one-step reaction from these abstracts, including all possibilities in the case of multistep reactions, was selected, giving a total of 582 reactions. The reactant and product molecules were encoded in fully expanded WLN and a printed index was

produced for the set of reactions using the Sheffield programs[4]; this index was searched manually. The CRDS file, which includes the set of reactions under consideration, was searched using programs written at Pfizer for that purpose. The appropriate volumes of Theilheimer were also searched, both by the manual coding system and by the keyword index. The purpose of this was to ascertain whether the keywording or codes would be useful in an specific situation where the structural description did not perform well.

A set of 18 queries was then constructed which was intended to represent the variety of reaction searches which a general purpose system should deal with; both general and specific queries were included. Because of the small size of the data base there were in general few examples of each reaction type; this is to be expected from the known distribution of reactions[9,10] and is not greatly deleterious to the qualitative evaluation attempted here. Each abstract in the data base was examined to determine those reactions relevant to each of the queries. This provided the ideal response sets against which the performance of the systems could be measured.

One example of the searching procedures is given here by way of illustration The example shown in Fig.1c involves the replacement of an aldehyde by a cyano group.

Relevant reactions will be analysed by the WLN algorithms[4] as

$$-CHO \longrightarrow -CN$$

or

$$-CH_2CHO \longrightarrow -CH_2CN$$

where the groups may be attached both to rings and to acyclic substructures. Therefore the printed index was scanned under the reactant site fragments *VH,/VH, *1VH and /1VH[4] and then the possible reactions checked by consideration of the reaction site notations.

In CRDS the (formal) breaking of C—H and C=O bonds, and formation of a C≡N bond were encoded. The codes for aldehyde in the reactant and cyano group in the product were also included. The search output was the Theilheimer abstract numbers.

3. Results

Seven relatively simple functional group interchange reactions were first considered. These are shown in Figure 1. Both the WLN systems and CRDS worked well on these examples. The first five examples were searched straightforwardly and all relevant answers found by both systems. In examples f and g, one possibly relevant reaction was missed by the WLN: this reaction was of the form

$$R{-}COCHBr \longrightarrow R{-}COCHD$$

Because the WLN analysis algorithms produce the most detailed description possible of the reaction site[4], the analysis was

$$-COCHBr \longrightarrow -COCHD$$

rather than the more general

$$-CHBr \longrightarrow -CHD$$

This is an example of the precision of the WLN approach. For a more general acyclic search, it is at present necessary to consider possible subsections of the reaction site character strings. Note that the reactions f and g require different coding in CRDS since the latter requires specificatic of both the bromine and the deuterium: the WLN searches are identical. The manual Theilheimer coding also proved reasonably efficient for these simple queries, although it involved a good deal of manual scanning.

A more specific query is shown in Figure 2 which involves a consideration of the reactant/product structures. In the case of CRDS more specific Ringcoding than in the general case was needed; with WLN, scanning of the printed notations was sufficient - in a computerised system a stringsearch could be used. Both examples were straightforwardly searched in the two systems and the relevant reactions found.

Two elimination reactions, one with greater structural specificity, were tested (Figure 3). Both systems found the relevant answers for the more general case, and 3b was found by product structure search.

Two somewhat more complex reactions were then examined, as shown in Figure 4. For 4a, the addition of methyls to an unspecified substrate,

the lack of information made it impossible to code any reactant or product structure for the CRDS. The large output resulting from use of the rather general reaction code included the relevant reactions. The relevant answers were found in WLN by scanning the full notations of those reactions involving the gain of two methyl groups. In example 4b a WLN search was possible by looking through examples of formation of all appropriate heterocyclic rings, which retrieved the relevant examples. CRDS produced the relevant reactions, but with many spurious answers due to the ill-defined query.

Two ring reaction queries were considered (Figure 5). The specific formation of a C—C bond within a defined heterocyclic ring in 5a presented no problems to either of the systems, both of which produced the relevant reactions from a straightforward search. The same was found with 5b, where the presence of a carbonyl linkage in a ring in both reactant and product gives structural specificity. It is worth noting that in both these cases a search in the Theilheimer volumes via the reaction coding would be highly inefficient, since all the sections corresponding to formation or breaking of C—C bonds would have to be scanned. The CRDS system allows specification of reactant and product structure, while in the WLN system the reaction site fragments include the whole ring formed or broken.

Finally, four more general queries were selected, as shown in Figure 6. These in general caused the greatest problems to the systems.

In the first three examples the structural environment is ill-defined. For all of these queries only acyclic WLN searches were made since in cyclic structures the reaction site fragments would comprise the whole monocycles involved in the change and each of these would have to be separately searched. No coding at all can be produced for a CRDS search for 6b, while 6a can only be coded for "formation of a C—H bond", giving rise to many errors.

Keywording for general concepts, e.g., "hydrogenation","double bond migration", seems a more feasible way of dealing with general concepts of this sort. Thus in example 6c, the relevant reactions may be readily found

from the index to the Theilheimer volumes, under the heading "ketones from allenes", emphasising the value of keywording for concepts of this sort.

For query 6d, the relevant answers can be found from the WLN only by scanning all appropriate heterocyclic rings formed, an impractical procedure for a large data base; similarly the CRDS search produces a very large output, because of the generality of the structure change. The relevant reactions are readily found from the Theilheimer volumes by the keyword phrase "replacement of oxygen, cyclic, by nitrogen/sulphur, cyclic".

Several of the queries were retested, using only the reaction (bond change) coding of CRDS, without reactant or product structures. In all cases a very large number of answers, many erroneous, resulted from the general coding. The specification of reactant and/or product structures by Ringcode is obviously an essential component of this system in a practical situation. The erroneous output from CRDS used in this way represents, as might be expected, a wide variety of reactions involving the same type of bonds broken or formed. When CRDS is used with structures specified relatively few erroneous results appear. These are usually due to a bond change in a different part of the structure; thus the reaction shown in Figure 7a was retrieved as an answer to 6c and that in 7b as an answer to 1d.

It is difficult to make a direct comparison with likely errors in the WLN system, where the search was carried out by scanning a printed index. It is evident, from the number of occurrences of the various fragment keys from the WLN analysis, that some form of structure search may be necessary to limit the output. However, increased precision in searches may be obtained from the fact that a structured feature may be specified as being actually involved in the reaction, rather than merely being present in one of the reacting molecules, by its presence in the reaction site notation.

Discussion

The most immediate impression gained from the results of this comparison is the great similarity between the performance of the two systems. In

general, reactions occurring in well-defined structural environments may be searched efficiently by either system, whereas more generally stated queries are poorly dealt with. There are, nonetheless, distinct differences, as will be noted below.

An evaluation of this sort makes clear the large extent to which a reaction information system requires a structure search capability. The CRDS system requires the specific coding of reactants and products to reduce output to a manageable level. The WLN system to some extent incorporates structural information by including larger fragments in its reaction site analysis, but may still require some examination of reactant and product structures for maximum effectiveness. In many cases, however, the reaction site notations are sufficient to characterise the change. In a computerised system based on WLN a substructure search procedure would be required, operating on the reaction site and/or the full reactant and product notations. The relative merits of Ringcode and WLN for substructure search would then have to be considered in a comparison of these reaction systems.[11]

The inclusion of considerable structural information in the reaction site notation often enables the WLN system to give a more precise analysis than CRDS. This is exemplified by the search for reaction 1f, where the presence of a ketone adjacent to the reaction site gave a different analysis, and in the ring formation and closure reactions where the monocycles involved were delineated both by the fragments and the reaction site notation. This is a very powerful feature of this type of analysis. Frequently reaction queries are specified in just this way, i.e., in terms of precise groups and ring systems, and an analysis based on WLN gives a rapid and reliable result. This is due to the extent to which such analyses retain the ability of the notation to describe structures in accordance with chemical intuition. In other cases, however, the two types of analysis are comparable.

Both systems are currently poorly equipped for handling the more general queries, i.e. those involving particular structural modifications in a variety

of environments. For a manually-indexed file, these problems could be alleviated by the use of intellectually assigned keywords similar to those used in the indexes of Theilheimer; this has been proposed by Derwent for the on-line version of CRDS. In the case of the WLN system, generic search capabilities could be obtained algorithmically both from the reaction site and parent compound notations.[12]

In summarising the performance of the two systems it is useful to consider the various access points to reaction information provided by the systems, and their appropriateness to the several types of reaction query likely to be encountered.

The two systems, as noted above, allow for approaches to reactions at different levels of structural specificity. CRDS allows searching directly for bonds formed and broken: in the WLN system this can only be achieved indirectly, by considering the possible reaction site changes brought about by a given bond change. The WLN system allows for a direct search on reactio sites: in CRDS an indirect search, combining bonds changed with structural features present or not present in reactant and product, is required.

The WLN approach gives three levels of structural description[4]: reaction fragments, reaction site notations, and parent structure WLNs. CRDS allows two levels: bond changes, and bond changes plus structural features of reactant and product (which may or may not form part of the reaction site).

It is helpful, accepting some degree of over-simplification, to consider possible reaction enquiries as falling into three classes: structural concept related, reaction site related, and bond related.

Concept related questions are typified by the more general test queries above. They are expressed as structural concepts, as exemplified by the reactions of Figure 6 , but are not restricted to anything other than a very general structural environment. Such questions are poorly dealt with by the structural reaction descriptions of both systems, and some form of concept indexing is desirable.

Reaction site related queries involve specification of the bond changes, with sufficient information on surrounding atoms to give a description of the reaction in chemically significant units, functional groups, ring systems etc. It may well be that this type of query will predominate for a general organic reaction information service. These queries are dealt with by using the reaction site information from WLN, or the bond change with reactant and product structures in CRDS. As noted above, both systems dealt effectively with test questions of this sort, with the WLN system having some advantages.

Bond related queries involve specification of a bond or bonds broken or formed, without full specification of the reaction site change: such information could be particularly useful for synthetic planning. Direct access at the bond level would be a valuable component of a comprehensive reaction information system: this could be provided by manual indexing, as in CRDS, or by algorithmic means.[13,14]

Any of these three types of query may involve specification of structural features of the reacting molecules, not involved in the reaction site. This may be achieved in both systems, using the substructure searching capabilities of WLN and Ringcode respectively.

A comparison of ease of use of the two systems would not be entirely meaningful, since much of the complexity of use of CRDS is due to structure searching via Ringcode. The corresponding computerised WLN searching was not undertaken: consideration of this factor again brings up the comparison of WLN and Ringcode as structure representation.

A WLN system is inherently more flexible than a fragmentation based system, since it allows production of printed indexes with whole structure representations provided as well as computer searches. This means that a WLN reaction system can provide both hard-copy output and printed desk-top tools which, given a working knowledge of WLN, could be used by the bench chemist. A fragmentation code system can only be efficiently used via computer, unless it relies upon a restricted coding such as the Theilheimer

code. It may be that a printed index WLN reaction analysis system would be best used as an aid to immediate synthetic problems, perhaps with relatively small files. In this way full advantage could be taken of its ability to rapidly answer precise questions of the kind often encountered in day-to-day synthetic work. A useful application would be reaction indexing of internal data banks, where structures are already coded in WLN and where existing WLN handling programs could be utilised. This would give a reaction searching capability entirely compatible with in-house structure searching. For larger files, a computerised search system would probably be required.

Conclusions

The two systems, based on WLN analysis and on bond change descriptors with Ringcode, were both found to deal effectively with queries defined in terms of reaction site change, involving functional groups, ring systems, etc. Such queries may well predominate in general purpose reaction information systems. The WLN system provided greater precision in some cases, due to the varying levels of structural representation provided. For some questions the bond change information in CRDS is valuable: this may be particularly useful for synthetic planning. Both systems perform poorly with concept related queries, where there is no specific reaction site indication. They both require some form of concept indexing for full overall effectiveness. A WLN-based system may be valuable in providing printed indexes of reactions.

Acknowledgements

We thank Mr. D. A. Faulkner for valuable discussion and assistance with WLN encoding, and Dr. A. Finch of Derwent Publications for advice on CRDS usage. Peter Willett thanks the Department of Education and Science for an Information Science Research Studentship.

References

1. Valls, J. Reaction documentation in Wipke, W. T., Heller, S. R., Feldman, R. J. and Hyde, E. eds., Computer representation and Manipulation of Chemical Information, John Wiley, New York, 1974, pp.

2. Valls, J. and Schier, O. Chemical reaction indexing in Ash, J. E. and Hyde, E. eds., Chemical Information Systems, Chichester, Ellis Horwood, 1975, pp.

3. Willett, P. The automatic analysis of chemical reaction data. Information Scientist, 11(4), 1974, 125-135.

4. Lynch, M. F. and Willett, P. The production of machine-readable descriptions of chemical reactions using Wiswesser Line Notations, J. Chem. Inf. Comp. Sci., (in press).

5. Derwent Publications Ltd., Rochdale House, 128 Theobalds Road, London, WC1.

6. Schier, O., Nübling, W., Steidle, W., and Valls, J. A System for the Documentation of Chemical Reactions, Angew. Chem. Int., 9(8), 1970, 599-604.

7. Theilheimer, W. Synthetic Methods of Organic Synthesis, vol.1; Basel, New York, 1946.

8. Nübling, W. and Steidle, W. The "Dokumentationsring der chemisch-pharmazeutischen Industrie"; aims and methods, Angew. Chem. Int., 9(5), 1970, 596-598.

9. Garagnarni, E. and Bart, J. C. J. Organic reaction schemes and general reaction - matrix types, III. A quantitative analysis, Z. Naturforsch, 32B, 1977, 465-468.

10. Lynch, M. F., Nunn, P. R. and Radcliffe, J. Final Report to the British Library, Research and Development Department on the project "Development of and assessment of an automatic system for analysing chemical reactions' BLR&D report 5236, London, 1975.

11. Sasamoto, M , Kubota, T., Hamano, T., Shinba, T. and Nakai, M. A qualitative comparison of Wiswesser Line Notation with Ringcode, J. Chem. Doc., 13(4), 1973, 206-211.

12. Granito, C. E., Becker, G. T., Roberts, S., Wiswesser, W. J. and Windlinx, K. J. Computer-generated substructure codes (bit screens), J. Chem. Doc., 11(2), 1971, 106-110.

13. Vleduts, G. E. Development of a combined WLN/CTR multilevel approach to the algorithmic analysis of chemical reactions in view of their indexing, BLR&D report 5399, London, 1977.

14. McGregor, J. J. and Willett, P. unpublished results.

(a) $R-NO_2 \longrightarrow R-NH_2$

(b) $R-O\overset{\overset{O}{\|}}{C}CH_3 \longrightarrow R-OH$

(c) $R-\overset{\overset{O}{\|}}{C}H \longrightarrow R-CN$

(d) $R-H \longrightarrow R-NO_2$

(e) $\begin{matrix} R_1 \\ \diagdown \\ C=O \\ \diagup \\ R_2 \end{matrix} \longrightarrow \begin{matrix} R_1 \\ \diagdown \\ CH-OH \\ \diagup \\ R_2 \end{matrix}$

(f) $R-X \longrightarrow R-H$ or $R-D$

(g) $R-Br \longrightarrow R-D$

In all of the figures R, R_1, R_2 etc. represent any unchanged feature.

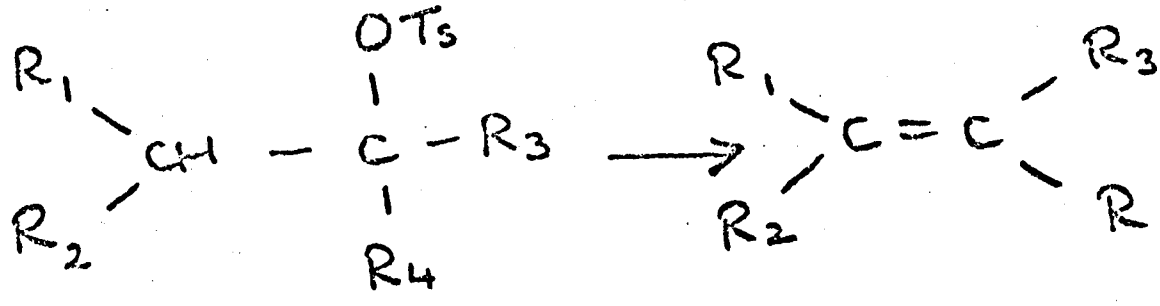
In this case R_1 and R_2 may not be part of the same ring and X represents a halog

Figure 1.

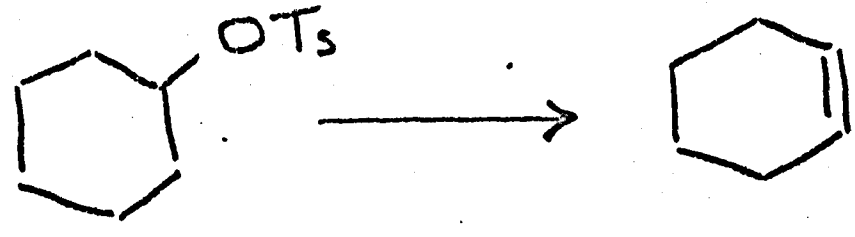
A ketonic group must remain unreduced in R

Figure 2.

(a)

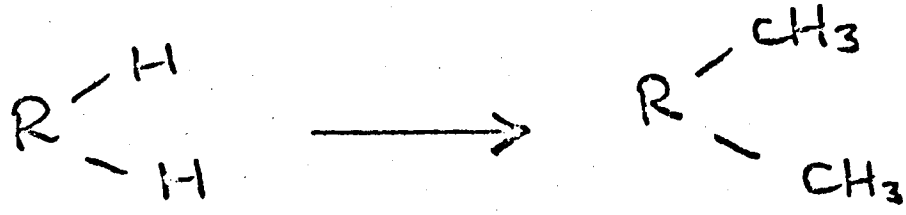


(b)



Ts is para toluene sulpfonyl

Figure 3.

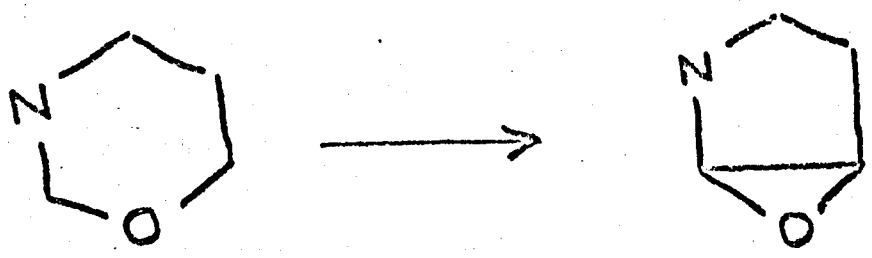
(a) 

(b) $R_1NH-X-NHR_2$ ⟶ a 5 or 6 membered ring with meta nitrogen ato

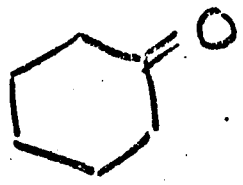
X is any carbon atom

Figure 4.

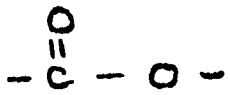
(a)



(b)



opened whilst a ring containing



is unchanged

Figure 5.

(a) $-CH=CH- \longrightarrow -CH_2-CH_2-$

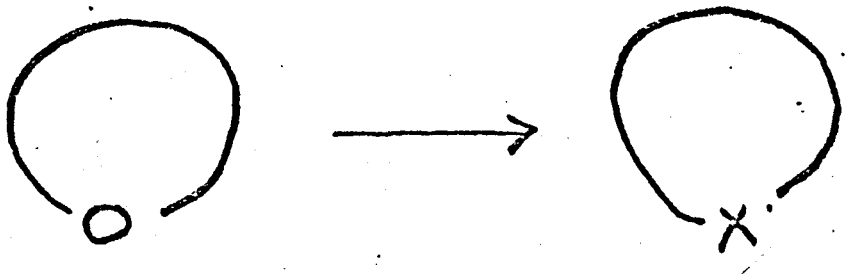
(b) $R_1=CH-CH_2-CH=R_2$

$\longrightarrow$

$R_1-CH=CH-CH=R_2$

(c) $-CH=C=CH- \longrightarrow -CH_2-\overset{\overset{O}{\|}}{C}-CH_2-$

The three reactions above may occur in any environment

(d) 

The circle represents any sort of ring and X represents nitrogen or sulphur

Figure 6.

(a) $R_1-C\equiv C=C-R_2 \longrightarrow R_1-\overset{\overset{O}{\|}}{C}-CH_2-CH_2-R_2$

(b) $X-N=N-\langle O \rangle-NO_2$

$\longrightarrow$

$\langle O \rangle-N=N-\langle O \rangle-NO_2$

X is any halogen

Figure 7.