(2) Dittmar, P. G.; Stobaugh, R. E.; Wilson, C. E. "The Chemical Abstracts Service Chemical Registry System. I. General Design". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111-121.
(3) Wigington, R. L. "Machine Methods for Accessing Chemical Abstracts Service Information". In "Proceedings of IBM Symposium on Computers and Chemistry"; IBM Data Processing Division: White Plains, NY, 1969.
(4) Richman, S.; Hazard, G. F., Jr.; Kalikow, A. K. "The Drug Research and Development Chemical Information System of NCI's Developmental Therapeutics Program". In "Retrieval of Medicinal Chemical Information". *ACS Symp. Seri.* **1978**, *No. 84*.
(5) Schenk, H. R.; Wegmuller, F. "Substructure Search by Means of the Chemical Abstracts Service Chemical Registry II System". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 153-161.
(6) Graf, W.; Kaindl, H. K.; Kniess, H.; Schmidt, B.; Warszawski, R. "Substructure Retrieval by Means of the BASIC Fragment Search Dictionary Based on the Chemical Abstracts Service Chemical Registry III System". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 51-55.
(7) Graf, W.; Kaindl, H. K.; Kniess, H.; Warszawski, R. "The Third BASIC Fragment Dictionary". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 177-181.
(8) Farmer, N. A. "The Proposed Chemical Abstracts Service's Substructure Search System". In "Proceedings of the Technical Information Retrieval Committee of the Manufacturing Chemists Association"; Arlington, VA, Aug 1977; McNulty, P. J., Smith, R. B., Eds.; Manufacturing Chemists Association: Washington, DC, 1977.
(9) Zeidner, C. R.; Amoss, J. O.; Haines, R. C. "The CAS ONLINE Architecture for Substructure Searching". In "Proceedings of the 3rd National Online Meeting"; Learned Information, Inc.: Medford, NJ, 1982; pp 575-586.
(10) Hagadone, T. R.; Howe, W. J. "Molecular Substructure Searching: Minicomputer-Based Query Execution". *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 182-186.
(11) Blake, J. E.; Farmer, N. A.; Haines, R. C. "An Interactive Computer Graphics System for Processing Chemical Structure Diagrams". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 223-228.
(12) Dittmar, P. G.; Mockus, J.; Couvreur, K. M. "An Algorithmic Computer Graphics Program for Generating Chemical Structure Diagrams". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 186-192.
(13) Feldman, A.; Hodes, L. "An Efficient Design for Chemical Structure Searching. I. The Screens". *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 147-152.
(14) Lynch, M. F. "Screening Large Chemical Files". In "Chemical Information Systems"; Ellis Horwood: Chichester, 1975.
(15) Dunn, R. G.; Fisanick, W.; Zamora, A. "A Chemical Substructure Search System Based on Chemical Abstracts Index Nomenclature". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 212-218.
(16) Stobaugh, R. E. "The Chemical Abstracts Service Chemical Registry System. VI. Substance-Related Statistics". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 76-82.
(17) Mockus, J.; Stobaugh, R. E. "The Chemical Abstracts Service Chemical Registry System. VII. Tautomerism and Alternating Bonds". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 18-22.
(18) Blackwood, J. E.; Elliott, P. M.; Stobaugh, R. E.; Watson, C. E. "The Chemical Abstracts Service Chemical Registry System. III. Stereochemistry". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 3-8.

DARC Substructure Search System: A New Approach to Chemical Information†

ROGER ATTIAS

Association pour la Recherche et le Développement en Informatique Chimique (ARDIC), 25 Rue Jussieu, 75005 Paris, France

Received November 29, 1982

The efficiency of a chemical information system depends upon information parameters retained by the language used to describe compounds. Structural-based languages provide a specific approach to chemical problems. The DARC system allows a coherent approach to substructure search, structure–activity correlation, and computer-aided design by defining relationships between the notions of substructure, structure, and family of structures. The substructure search is based on the concept of fuzziness: it is expressed in terms of subgraph isomorphism between a set of fuzzy graphs and a set of graphs. The file to be searched is processed by an automat which generates multilevel fuzzy graphs corresponding to local descriptions of the defined structures. The DARC descriptions of these graphs are stored in a tree structure. The same process is applied to the fuzzy graph of a query. As a result of this approach, the user language is the natural language of the chemist: free drawing of the substructural diagram with no use of a dictionary for the search. The retrieved structures can be displayed on a graphic terminal. These principles have been applied to the full CAS Registry Structure File and have made possible, for the first time, on-line substructure searches on 5 million compounds (EURECAS). An automatic link to the textual data base (CA SEARCH) makes it possible to deal with both structural and textual aspects of the query.

INTRODUCTION

The DARC system and its role in French national computer science policy were presented in 1972 by Professor Jacques-Emile Dubois.[1] This system, developed since 1963,[2-6] places chemical compounds in their structural context and accounts for their local and global properties by using their topology as a starting point.[7,8]

Structural information is handled so as to achieve a coherent approach to the notion of substructure, structure, and family of structures (hyperstructures).

The substructure is perceived as a generalization of the notion of structure. The substructure search system, which is an application of the basic principles, constitutes the first and necessary step forward in handling chemical problems. One aim of this system was to make possible access to Chemical Abstracts Service (CAS) products, not only by texts but also by structures through a structural user language reflecting the thought of the chemist and enlarging his field of investigation. To achieve these goals, our constraint was to perform a purely topological approach, avoiding, whenever possible, any type of global fragmentation. The first tests were on samples;[8,9] the difficulties then perceived as to volume and response time were gradually solved and led to on-line search in 1978 of CBAC and in 1980 of EURECAS (commercialized in Feb 1981).

In this paper we introduce the different steps of our research, its context, its results, the general methodology which has guided our approach, and its place in the evolvement of structural languages: the concepts and the technical aspects will be more fully developed in a series of papers in which, in particular, will be discussed the important contributions of

† Presented at the 182nd National Meeting of the American Chemical Society, New York, August 1981.

0095-2338/83/1623-0102$01.50/0

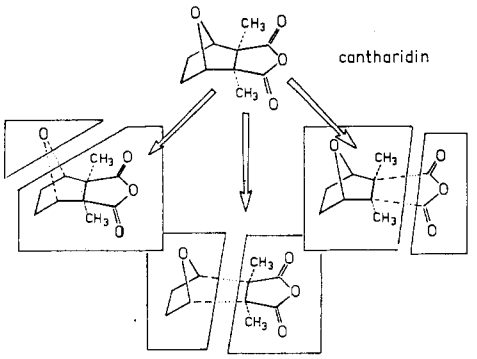

Figure 1. To carry out a synthesis, the chemist tries different approaches by isolating certain substructures of the target structure. These substructures can be described precisely only by their structural diagram.

other screening techniques in this field of substructure search (Lynch, Lefkowitz, Feldman, Hodes, ...).

After having stressed the basic role of the substructure in chemistry and its expression, we introduce the fundamental principles of the DARC system and explain how they are used to solve the problem of substructure search. The resulting query language is structural, and we analyze its possibilities. Finally, we introduce the achievements which led to the first on-line service of the complete CAS Registry Structure File, while emphasizing the potential of this huge volume of structural information and the advantage of computerized coupling of structural and textual queries in CBAC and EURECAS products.

I. SUBSTRUCTURE AND LANGUAGE

The radical theory—groups of atoms behaving as a whole and remaining unchanged during certain reactions—constituted an essential step in developing a systematic nomenclature of chemical compounds and was the first approach to the substructure notion. This notion has, since then, been the topic of much research endeavoring to reply to the chemist's concerns: correlation, drug design, patent protection, elucidation, synthesis pathways, documentation, ... For each of these areas, one step of the chemist's procedure is to search for structures bearing common structural characteristics, varied and complex, for which the only possible expression is a specific structural diagram.

For example, the search for a starting compound for cantharidin synthesis required successive approaches based on the definition of substructures extracted from the target structure (Figure 1). Another example is the search to attribute properties to a substructure yet to be determined and which, by definition of the problem, corresponds to no function yet identified. The potential of a chemical information system depends on the perception of the chemical compound and, therefore, on the language which describes it. The name given to a substance[11] has become more and more precise with improved knowledge of the composition and of the organization of the reality it describes: a name linked to appearance (consistence, color, savor, odor), a name linked to properties, a name linked to an imperfectly known composition, then a molecular formula, and finally a systematic name.

A systematic name uniquely describes the structural diagram of a compound by identifying the ring systems, the usual chemical functions, and the adherence to a series. Codification of these elements resulted in fragmentary codes, a compact but ambiguous description of chemical compounds. The limited descriptive quality of the fragments makes it impossible to grasp unusual classes of compounds in a classical taxonomy. Linear notations (more complex) break up fragments by limiting the number of codes used: they remain ambiguous in certain cases. In order for specialists to handle these notations, certain constraints are necessary which limit their descriptive potential and their optimal use by computer.

The evolvement of descriptive techniques is analogous to that of knowledge of a compound: global descriptions of structural fragments are replaced by a description in terms of the component elements of the structure. This type of description is achieved by topological codes which preserve the integrity of a chemical compound's structural information.

The language used to describe a structure plays an essential role in the substructure approach: preciseness of structure description determines flexibility in defining substructures to be processed in a chemical system and therefore the features of user language. Parameters of the substructures extracted from structures are generic expressions of some structure parameters, the highest specificity being necessarily a specificity from the structure. The preciseness of CAS topological code allows structural processing for a vast collection of compounds.

One aim of the DARC approach of structural languages is to elucidate topological relationships between substructures, structures, and ordered populations of structures.

II. BASIC PRINCIPLES OF DARC SUBSTRUCTURE SEARCH SYSTEM

A structure is progressively described by a succession of concentric constant environments around a focus. Substructures are defined by means of the same local ordering rules. In order to handle a larger variety of structural indeterminations, we define conceptual entities introducing the notion of a fuzzy structure whose generic description allows a progressive and hierarchically organized description of real entities.

(1) Reminder of the Structure Description Method. The structure is not considered as an isolated entity; it is situated in its structural context through a progressive description. An origin is selected in a unique way. The successive concentric layers of atoms around this origin are alternatively labeled A and B. The atoms filiated to the origin and belonging to the first A and B layers are considered as a whole and are ordered by a function which attributes to each atom a total order labeling A_i and B_{ij} on the basis of atom connectivities, atom values, and bond values. They constitute the first ELCO (Environment which is Limited, Concentric, and Ordered). Recursively, each B atom is considered in turn as the origin of the ELCO comprising the filiated atoms on the next A and B layers.

Ordering rules define the ELCO's relative order, thus inducing a total order on the whole of the structure, on the basis of progressive local ordering of canonical structure subsets.

The DEL (Description by Environment which is Limited) is the resulting unambiguous linear code of the structure describing the ELCO's and their relative position in the structure.

(2) The FREL. The order carried out on a structural entity depends on this entity considered as a whole. The ELCO has its full meaning only in the context of the structure it describes. In order to fully apprehend local environments of a structure, we use another DARC basic substructural unit, the FREL (Fragment Reduced to an Environment which is Limited), which describes two concentric layers of atoms around a focus. The FREL's nodes are ordered by the ELCO ordering function, structures and substructures being thus homogeneously described. A specific and unique ordering of structure fragments, independent of global ordering of the structure, is achieved.

The FRELs are generated algorithmically from structures according to topological criteria and reflect locally the relative positions of cyclic and acyclic elements; they constitute an open

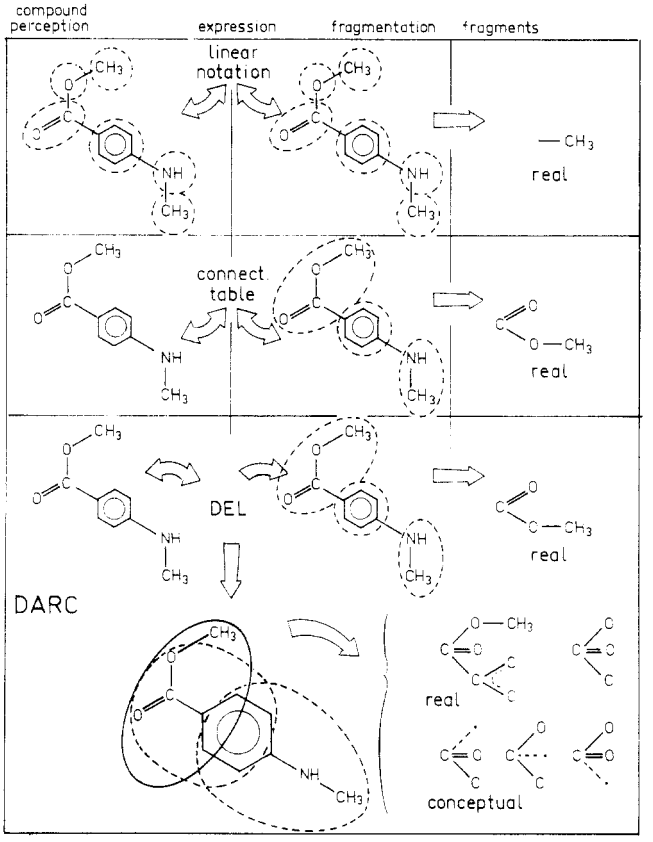

Figure 2. The fragments used in linear notations are based on recognition of chemical functions or elementary entities. Nomenclature fragments are constituted of ring systems, chains, and functional groups. These two types of fragmentation have been defined to fulfill the function of naming and representing a structure. FRELs are canonically defined in order to locally apprehend a structure; they are overlapping topological fragments precisely describing two layers of atoms around a focus. Conceptual entities comprise undetermined features and are the expression of generic local description. At the bottom of the figure, examples of fragments corresponding to FRELs are circled. Real and conceptual entities, corresponding to FRELs and canonical fuzzy FRELs, are detailed for the solid-line circled FREL.

set, characteristic of the corpus from which they stem.

Connectivity degrees determine the choice of foci. FRELs generated from high connectivity degree foci describe the topologically richest parts of the structure, while unbranched parts of rings and linear chains are well described by FRELs centered on atoms whose connectivity degree is 2. The module described by a FREL is broad enough to characterize a certain specificity of the compound. Certain FRELs can characterize classical chemical functions, classical functions in a given environment, or a group of atoms corresponding to no preexisting classification and possessing certain properties.

The set of FRELs of a structure constitutes a description of this structure, aiming less at identifying a structure in a nonambiguous way than at replacing a global approach by one which proceeds by parts to tackle chemical problems. Redundancy of this description allows a varied apprehension of a site differently located in different neighboring environments. Overlapping limits the possibilities of combining FRELs to reconstitute a structure (Figure 2).

(3) The Fuzzy FREL. A substructure comprises undetermined structural features which also appear in the topological fragments extracted from it. We have defined categories of generic fragments by introducing a new concept: the fuzzy FREL.

A fuzzy FREL comprises one or more atoms, bonds, or connectivity degrees with an undetermined value (all possible values) or a generic value (list of possible values). Its expression is the generic expression of a set of FRELs. We call infra-FRELs the fuzzy FRELs bounded by the first row of atoms of a FREL. Canonical fuzzy FRELs with varying increasing specificity are represented in a tree structure. A FREL is thus described gradually by a set of hierarchically organized generic structures. The different specific FRELs belonging to the class represented by any fuzzy FREL can be identified by traversing the tree structure.

This handling of structural information constitutes a new approach, by allowing undetermined structural aspects to be fully expressed and exact specifications of fuzzy fragments to be subsequently processed. The vast combinatorial possibilities of structurally indeterminate fragments cannot be stored by anticipation; the solution proposed consists of elucidating each of them by a dynamic process based on a topochromatic classification.

III. APPLYING THE PRINCIPLES TO STRUCTURAL SEARCH

Unambiguous topological representation associates a unique chain of bytes to every structure; the equality of two codes compared byte to byte implies the identity of the structures they represent. The problem of searching a file for a structure is thus reduced to the problem of searching for a key within a set of keys and can be solved by classical efficient retrieval methods such as Hash coding. A method based on specific structural parameter calculation has had very satisfactory results.[12] In the DARC system, a hash function based on binary addition of constant successive byte strings has been defined on DELs. This function yields well-scattered values and is particularly easy to compute since the argument is an alphanumeric code.

Substructure searching, related to the subgraph isomorphism problem, is a very different type of problem. Labeling is different in a substructure and in its impact on a structure; moreover, substructure indeterminate features have to be elucidated in the searched structures. Comparison between a structure and a substructure, which is therefore not reducible to a global codification process, is performed by atom-by-atom and bond-by-bond comparisons. Different optimizations tend to reduce the large combinatorial possibilities, but this sequential iterative process remains slow. The computer time required is directly proportional to the number of structures handled and becomes prohibitive when this number is great, in particular for an on-line search.

The general principle of any substructure search method is to select, by rapid but approximate screening techniques, a limited set of structures including the relevant ones. Direct comparison is only carried out in this limited file. A system's potential is linked to the capacity of its screening system to describe with precision the determined structural characteristics and to handle the indeterminate aspects of the structural question.

The fundamental principle of the DARC substructure search system is to replace the subgraph isomorphism global approach by a subgraph isomorphism local approach. The fuzzy FRELs extracted from the substructure are compared to the FRELs of the searched structures. At this local level, labelings are again different within each of the topological fragments to be matched (Figure 3), but the combinatorial possibilities are considerably reduced, and this is mainly due to the following. (a) The focus, starting atom of the comparison process, is uniquely preselected in the two structural entities to be matched. (b) The structural entities handled are canonically generated and correspond to a constant limited environment. (c) The FRELs, as a preliminary, are extracted from searched structures and organized in a tree structure whose leaves contain preselected fuzzy FRELs in such a way

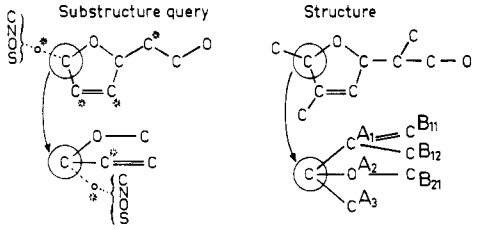

Figure 3. Example of a fuzzy FREL extracted from a substructure comprising indeterminate bond value (dashed line), atom value (list), and connectivity degree values (asterisks). The corresponding FREL extracted from a relevant structure has a different labeling. The subgraph isomorphism between the fuzzy FREL and the FREL is elucidated by traversing a tree of hierarchized canonical fuzzy FRELs.

as to anticipate and classify the indeterminations to be solved.

Substructure search is carried out by the following operations: generation of fuzzy and determined FRELs from the query, representation of each of these FRELs in the form of structural access keys and structural constraints, comparison of the query FRELs and those of the searched file by traversing the formerly defined tree, identification of relevant compounds for each query FREL in order to obtain the reduced file of possible answer structures, and atom-by-atom and bond-by-bond comparison of the substructure query and the reduced structure file to preserve only relevant answers, and all of these.

This method makes it possible to search for substructures in vast files (several million structures) with very short response times, compatible with on-line search. This is mainly due to the discriminatory nature of the FRELs in their capacity as structural screens (the slow process of direct comparison of the query is applied only to a limited number of structures, often close to the number of relevant structures) and the rapid comparison of the query FRELs with the hierarchically organized FRELs of the structure to be searched.

IV. QUERY LANGUAGE

The development of specifically structural instruments eliminates, for the chemist or the information specialist, problems of structure and substructure codification: these become an internal computer problem and are solved by the appropriate choice of the order defined on the entities handled. Substructure queries are directly put to the system by their structural diagrams, in a conventional structural representation, which leads to language simplicity, to exhaustive answers, and, above all, to the introduction of new search possibilities.

(1) Language Possibilities. The originality of a language for a chemical information system lies in its capacity to handle structural entities. The actual query submitted to a system is a translation into the query system language from the initial user question: after analysis, this question is adapted, and sometimes modified, to fulfill the search system requirements and capabilities. The fundamental topic of a language development is to provide the closest equivalence in meaning between the user question and the constructed system query. Efficiency is achieved only if formulating the system query is easy and mastering its significance is simple. The choice of a language induces search techniques and strategies which, in turn, set the upper limit of language optimization by defining the set of all potential searchable queries. After all searchable queries can be easily formulated through the language, the next improvement is addition of new search system features to enlarge this set.

The difficulty of formulating a structural query in terms of nomenclature has been fully discussed.[13,14] Lack of precision limits the possibilities; certain nomenclature search terms constituting the query could themselves be considered as part of the answer.

Development of substructure notion in the DARC substructure search system resulted in a user language allowing the formulation of a query by means of a structural diagram, where each of its component elements (atoms and bonds) is specified; indeterminations of atom values, bond values, and atom connectivities are directly indicated on the diagram by means of commands.

The use of short cuts is an option for quick formulation of a given set of bonded atoms; they are automatically expanded for query processing and do not modify the search. User questions containing features whose expression is not strictly described through a structural diagram, e.g., chains of varying length, ring attachment at unknown positions, etc., can be transformed into a Boolean expression of system queries.

The search is limited to queries comprising at least two consecutive bonds and their related atoms with a defined value.

Only structures corresponding to the exact query specifications (lists of atom values, bond values, and atom connectivities) will be retrieved. Two important consequences are as follows: for one thing, every structural aspect of the query is fully mastered by the user, and, for another, the query structural representation must be established with respect to the conventional structural representation adopted for the data base.

The main capabilities of the language are as follows.

(a) The value attributed to a bond or to an atom can be a single value, can be a list of values defined by the user, or can be left undetermined by the user. For example, if a list of values is attributed to an atom of a query, any structure having, in the corresponding position, one of the values on the list is a possible answer.

(b) The number of attachments to an atom other than hydrogen can, for a given atom, be imposed on or belong to an interval specified by the user. Atom substitutions can thus be forbidden, or authorized, by controlling the number. If a substitution is authorized for a query atom, the answer structures include all structures for which this atom is not substituted and all structures where a hydrogen atom is substituted on this atom, whatever the nature of this substitution. It is precisely the examination of the nature of these substitutions which will provide the information desired by the user. This process remains possible when the saturated state in hydrogen does not correspond to a chemical reality: if a possibility of attachment on an atom is permitted when the query is formulated, it is the precise nature of these attachments, present in the data base, which is obtained in the answers. We should note an important consequence with regard to query capacities: if a structural question comprises chains whose extremities have topological indeterminations, such as we have just described, the system will provide the following as relevant answers: the structures where these chains are acyclic; the structures where these chains are attached by a bond, or another chain, thereby constituting a ring. It is also possible to eliminate one of these two categories of answers.

This possibility of formulating ring elements in terms of chains removes the limitations set by systems based on nomenclature or on classical fragmentation, in which a total dichotomy exists between cyclic and acyclic elements of a structure.

(c) A ring system search does not require a global description when formulating the query; in reply to a query describing only part of a ring system, all ring systems including this part will be given, with respect to the constraints specified by the user. For example, if a query comprises a ring, with eventual indeterminations as to the nature of the atoms and bonds, the answers obtained will include all those ring systems in which this ring participates, according to the topological indeterminations specified on the query sites (Figure 4).

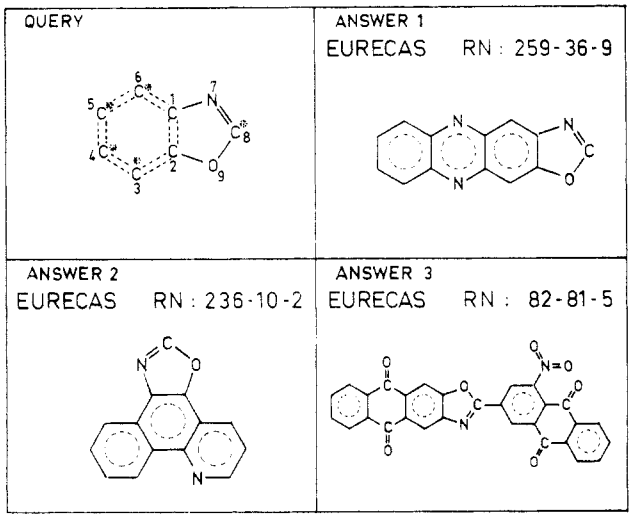

Figure 4. Example of a query comprising a ring system whose inclusion in larger ring systems is allowed. Search based on fuzzy FRELs and FREL generation leads to the answers containing the different relevant ring systems; each of them is a single classical fragment corresponding as to a nomenclature search term.

(d) A query may comprise several disconnected structural entities. These entities can be searched for in the same fragment (as defined by the CAS) or in different fragments of the same structure. This is an interesting possibility when relevant structures must comprise simultaneously two or several structural moieties whose structural relationships are poorly defined or indifferent. The AND function is thus achieved between substructures.

(2) The Language in Use. DARC substructure search functions are carried out by a set of simple commands hierarchically organized into three levels of commands. After a command on a given level is executed, the entire set of commands of this level becomes available. The command to end the use of a given level makes available the commands of the level above it. Without detailing the usage, we present here the different commands which complete a substructure search.

(a) Editing the Query. The first step for the user is to formulate the query in structural terms only, on paper, by using the indetermination possibilities. The query input begins with acquisition of the graph and then of the information associated with the graph edges and nodes (nature of the atoms, nature of the bonds, and indeterminations) (Figure 5).

Graph Input. This can be done either graphically or alphanumerically. Graphically, it is enough to draw the graph from a graphic terminal. The graph nodes are automatically numbered by the query interpreter. Alphanumerically, the input is done by linear notation of the graph (from an arbitrary numbering of the nodes). This method makes it possible to use alphanumerical terminals to carry out structural queries.

Chromaticity Input. For assignment of a value, nodes and edges are identified by the numbering assigned during graph input.

(b) Topological Screen Search. A single search command carries out the following functions: generating the structural screens of the query, searching the whole file by traversing the FREL tree structure, creating a file of possible answer structures.

(c) Iterative Search. An atom-by-atom and bond-by-bond comparison between the query substructure and the candidate answer structures leads to a file containing all relevant answers, and only these.

(d) Graphic Display of Answers. A command enables the user to display the two-dimensional structural diagram of the answers. Starting from the topological code, the program gradually generates the structure around a focus. Rings are

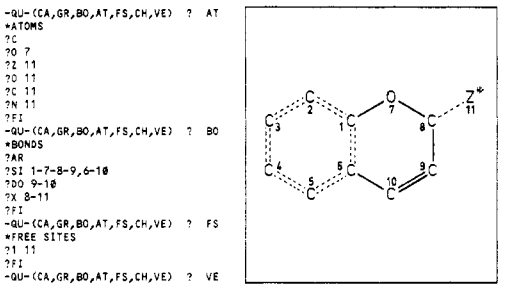

Figure 5. Example of the graphic input and check of a query. Atoms are automatically numbered by starting from the drawing of the graph on the terminal. Commands AT, BO, and FS make it possible to specify the values of atoms, the value of bonds, and the maximum number of attachments allowed on a site. Here, for example, three atom values are allowd for node 11. An example of alphanumerical input of this graph is 1–2–3–4–5–6–7–8–9–10–6,8–11. Numbering the nodes is arbitrary; a dash between two node numbers indicates the existence of a bond between these nodes; a comma is simply a separation.

detected automatically and displayed optimally.

The user language comprises a set of further commands (graphic and alphanumerical check of a question, modification of a query, Boolean operations, etc.) which we do not describe here.

Direct handling of structural entities simplifies the refinement and correction of queries which can be interactively modified. In chemical research, these operations can be carried out optimally by the end user whose steps can be guided by the series of interactions suggested by the answers obtained.

V. APPLICATION: EURECAS. FIRST ON-LINE STRUCTURAL SEARCH ON FIVE MILLION COMPOUNDS

(1) The System's Evolvement. Structural registration presented essential advantages in chemical data base management such as precision of identification, unique and unambiguous representation, and computer-generated nomenclature[16] and led CAS to develop a topological code. Since he had oriented work on the DARC system, from its inception, toward a homogeneous topological approach to problems of documentation, of structure–activity correlations, of elucidation, of drug design, and of computer-aided synthesis, Professor J. E. Dubois established continuous cooperation links with CAS and created ARDIC* (Association pour la Recherche et le Développement en Informatique Chimique), in charge of development.

An important preliminary stage was the study of the CAS topological code and the development of software to transcode it into an intermediary code, or pivot code, from which the DARC topological code is generated. Numerous exchanges took place with the CAS team, and this made possible the transcoding, first from registry II and then from registry III. The first tests of the DARC substructure search system were run on a sample of 10 000 structures from the CAS Registry Structure File (RSF) supplied by the Basel group in 1973. Statistical studies on the FRELs, studies of variation on hierarchically organized file volumes, and query performance studies were all carried out with precision, owing to the three samples of 20 000, 60 000, and 180 000 structures taken at random from a sample of 500 000 structures from among the 2 million then comprising the CAS RSF. We had previously transcoded into the DARC code these 500 000 structures provided by the CAS. The studies conducted on these samples resulted, in 1975, in the first version of the DARC substructure search system whose maintenance and development is assured by the ARDIC.

As of 1976, an operational version existed for the ITODYS (Institut de Topologie et de Dynamique des Systèmes) internal

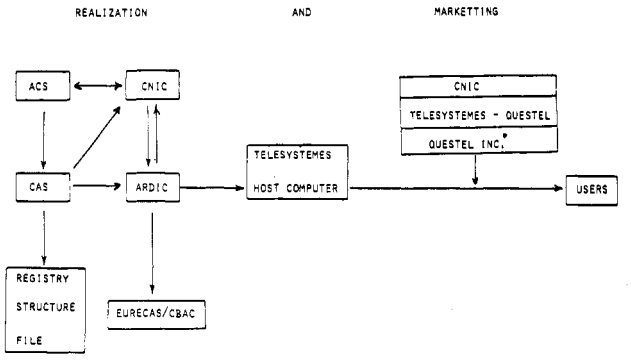

Figure 6. ARDIC, a nonprofit organization, supervised CBAC and EURECAS products and carried out the structural part through the use of DARC software applied to the CAS Registry Structure File supplied by the CNIC (license agreement CNIC-ACS 1978). The textual part was carried out by Tĕlĕsystèmes-Questel. The products are installed in its host computer in Valbonne (France) and are distributed for the CNIC. Questel Inc., Washington, DC, represents Tĕlĕsystèmes in the United States for distribution of EURECAS.

data banks on both DEC (PDP 11/35) and IBM (370/168) computers. In 1977, the system was demonstrated at the first London On-Line Meeting: the file searched was the CAS CBAC structural file which then comprised 125 000 structures. In 1978, the CNIC (Centre National de l'Information Chimique) became a partner of the CAS and entrusted to the ARDIC the task of implementing DARC substructure search software, applied to CAS structural products, on the CII/HB (IRIS 80 BP) national host computer in Valbonne (France), managed by Tĕlĕsystèmes-Questel. (Products have been subsequently transferred to an IBM computer.) Handling of the textual data (CBAC text files with abstracts and CA SEARCH) is conducted by Tĕlĕsystèmes-Questel. The first commercial version was demonstrated at the 1978 London On-Line Meeting on the CBAC product: 400 000 structures searched by the DARC system and 400 000 texts, whose abstracts were on line, searched by the Mistral software (now known as the Questel software). At the London On-Line Meeting in 1980, the whole CAS RSF[13] was searched on line for the first time.

Transcoding, as carried out on the whole RSF, preserves the CAS structural representation conventions. Optimizations were necessary in order to handle such volumes on line: on the level of search strategy, of data structure (the structure of hierarchically organized files, however, was not modified), and of user language.

The DARC search of CAS RSF coupled to the Questel search of CA SEARCH was offered as a public service under the name EURECAS, with world-wide access through the Transpac, Euronet, Tymnet, and Telenet networks (see Figure 6). A random sample of 1/100th of the RSF, known as MINICAS, allows for query checks and gives a good approximation of the volume of answers which would be obtained on the RSF.

A registry number–structure inverted file has also been developed. This allows queries using registry numbers as input.

(2) Combining Structural Information and Textual Information. Queries submitted to a chemical information system are often Boolean expressions comprising structural as well as textual elements.

An integrated process is proposed in the CBAC and EURECAS products to search structural and textual CAS data bases for these mixed queries: an automated link between DARC structural software and Questel textual software allows each type of expression constituting the query to be handled by the corresponding adapted software. Registry numbers corresponding to structural answers are transferred to the textual software, thereby providing corresponding bibliographic

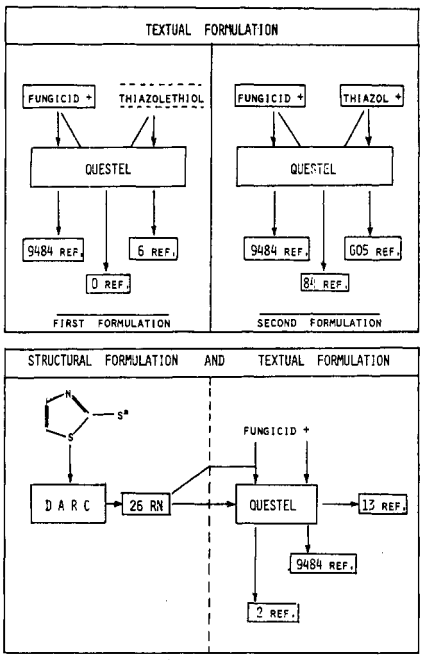

Figure 7. Different formulations of the question asked by a user: "What texts deal with the fungicide activity of thiazol family compounds?" (a) Textual formulation enumerates the different nomenclature search terms which seem to identify the thiazol family. This enumeration is incomplete and does not lead to any specific reference. (b) Textual formulation using the truncation technique is too broad and leads to 84 references of no interest to the user. (c) Structural formulation then enables him to precisely designate the structural family with which his search is concerned. He obtains 13 references (citing at least one of the 26 answer structures) whose intersection with the 9484 references citing the textual key word lead to two relevant references.

references and the search basis for a subsequent combined textual query. In order to accede to the exhaustive cited literature, by means of a structural approach, synonymity is maintained between the different registry numbers having been assigned to a same compound. These infrequent cases of registry number reassignments, due to re-registration of erroneously registered structures, are dealt with by processing replacing/replaced registry number records supplied by CAS. Figure 7 points out the gain in precision of such coupling compared to the exclusive use of textual software.

These combined softwares allow the advantages of classical documentation systems to be integrated in a graph-based processing system. Another approach of this integration would have been a global handling of the structural and textual information, for example, by the use of combined screens, comprising structural- and textual-type information. This approach can be interesting and an object of future study. However, in the context of the service offered, we have preferred the coupling solution for the following reasons. (a) One obtains processing specificity for data of a different nature (structural and textual); integration of the two would render the process more cumbersome but would shorten response time. Insofar as response times are compatible with an on-line system, this is not a technical necessity. (b) Structural search, on the complete CAS Registry Structure File, corresponds to a need and constitutes a service not necessarily linked to textual search. (c) When first put into use, the structural approach was a new one with which the user had to gradually become familiarized. It was important for him to also have at his disposal those search techniques to which he has previously been accustomed.

CONCLUSION

The structural concepts developed, resulting in a large-file

structural screening method adapted to generic structure manipulation, lead to accurate structural handling of chemical data bases. Extended possibilities of the structural language free the user from codification problems and allow a new field of investigation.

The CAS data base is the most complete computerized collection of chemical data and presents considerable potential. The flexible manipulation of generic structures brings out new possibilities for exploiting these topological data, of both practical and conceptual interest in the documentation and research fields in chemistry. For each of them, there is the need to represent and handle different types of structural entities, and it was important to evolve a system whose potential would not be limited by a reference base composed of restricted structural elements.

It is interesting to note that documentation has often been considered as a field apart from that of computer-aided design, which alone seemed worthy of the prestigious attribute of artificial intelligence. We feel that documentation, and in particular structural documentation, should not be reduced to processing files of rigid entities, independently from related problems; it represents a central element of design techniques insofar as it defines the expression and the representation of entities handled by the user on the one hand and by the computer on the other. Possible interactions with the user also make it an independent design tool.

The substructure search system is part of a set of computer-aided strategies developed within the framework of the DARC system.[17,18] Future developments will include growing aid to the user by interactive procedures integrated into the system.

ACKNOWLEDGMENT

I thank Professor J. E. Dubois for his encouragement with this paper and for the many fruitful discussions on this field of ordered topology in structural chemistry. We express our gratitude to the Chemical Abstracts Service staff for cooperative discussions and assistance with CAS topological code and structure representation conventions. These were very helpful in transcoding operations.

REFERENCES AND NOTES

(1) Dubois, J. E. "French National Policy for Chemical Information and the DARC System as a Potential Tool of This Policy". *J. Chem. Doc.* **1972**, *8*, 13. Presented at the 164th National Meeting of the American Chemical Society, New York, 1972.
(2) Dubois, J. E.; Laurent, D.; Viellard, H. *C. R. Hebd. Seances Acad. Sci., Ser. C* **1967**, *264C*, 348.
(3) Dubois, J. E.; Viellard, H. *Bull. Soc. Chim. Fr.* **1968**, 900, 905, 913.
(4) Dubois, J. E.; Panaye, A.; Viellard, H. *Bull. Soc. Chim. Fr.* **1973**, 1988, 1996.
(5) Dubois, J. E.; Alliot, M. J.; Panaye, A. *C. R. Hebd. Seances Acad. Sci., Ser. C* **1971**, *273C*, 224.
(6) Dubois, J. E.; Panaye, A.; Cayzerges, P. *C. R. Hebd. Seances Acad. Sci., Ser. C* **1980**, *290C*, 429, 441.
(7) Dubois, J. E.; Laurent, D.; Panaye, A.; Sobel, Y. *C. R. Hebd. Seances Acad. Sci., Ser. C* **1975**, *280C*, 851; **1975**, *281C*, 687.
(8) Dubois, J. E. "Structural Organic Thinking and Computer Assistance in Synthesis and Correlation". *Isr. J. Chem.* **1975**, *14*, 17.
(9) Dubois, J. E.; Bonnet, J. C.; Goldwaser, D.; Attias, R. "The DARC System—A Chemical Information System Based on the Topological Encoding of Chemical Compounds". Butter, W. E., Ed. "EURIM II, Proceedings"; 1976; p 135.
(10) Dubois, J. E. "DARC System in Chemistry". In "Computer Representation and Manipulation of Chemical Information"; Wipke, W. T., Heller, S., Feldmann, R., Hyde, E., Eds.; Wiley: New York, 1974.
(11) Goodson, A. L. "Graph-Based Chemical Nomenclature. 1. Historical Background and Discussion". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 167.
(12) Freeland, R. G.; Funk, S. A.; O'Korn, L. J.; Wilson, G. A. "The Chemical Abstracts Service Registry System. II. Augmented Connectivity Molecular Formula". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 94.
(13) Incompletely defined structures and coordination compounds were not taken into account in an initial stage in order to preserve homogeneity of handling and use.
(14) Fisanick, W.; Mitchell, L. D.; Scott, J. A.; Vander Stouw, G. G. "Substructure Searching of Computer-Readable CAS Ninth Collective Index Nomenclature Files". *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 73.
(15) Dunn, R. G.; Fisanick, W.; Zamore, A. "A Chemical Substructure Search System Based on Chemical Abstracts Index Nomenclature". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 212.
(16) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. "The Chemical Abstracts Service Registry System. I. General Design". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111.
(17) Mercier, C.; Sobel, Y.; Dubois, J. E. *Eur. J. Med. Chem.* **1981**, *16*, 473.
(18) Dubois, J. E. "Computer-Assisted Modelling of Reactions and Reactivity". *Pure Appl. Chem.* **1981**, *53*, 1313.