Substructure Systems: Concepts and Classifications

ROGER ATTIAS and JACQUES-EMILE DUBOIS*

Institut de Topologie et de Dynamique des Systemes de l'Universite Paris 7, associe au CNRS, 1 Rue Guy de la Brosse, 75005 Paris, France

Received January 5, 1989

The duality of the notion of substructure—whole vs part, element vs class—leads to varying levels of abstraction, each associated with different aspects of the evolving models for representation. The concepts underlying substructure definitions are discussed and are related to the problems of representation. Substructures are often goal-oriented, emphasizing some typical feature. The main approaches to substructure systems are reviewed in the light of various classifications, allowing representation of the same entity from different standpoints. A generic model, an extension of the graph-specific structure, provides a basic model which unifies structure and substructure and yields a formal tool for representation and evaluation.

INTRODUCTION

The concepts of *substructure* and *subsystem* are fundamental to reductionist theories and to knowledge theories. If a molecular paradigm allows us to deal with a number of pattern recognition problems, this is partly because it is based upon the "structure–substructure" duality with its associated properties. In the DARC system, we approached molecular data from the point of view of structural[1] and substructural[2,3] descriptions at various levels of size and generic character with filiation properties. The substructural unit used in DARC is the FREL (*F*ragments *R*educed to an *E*nvironment that is *L*imited), and for different types of FREL substructure, certain relationships between specific and generic knowledge were recently presented.[4]

Most applications in chemistry deal, in some way, with substructures, and these seem to be defined in very diverse ways. Definitions are often goal-oriented and emphasize some typical feature. The concept of substructure is so important in chemistry that we feel it worthwhile, aside from recognized molecular structure systems and their current applications, to consider the general nature of structure and substructure representation by means of graph theory, posing the questions: How can features common to different systems be defined? Is there a unique concept underlying definitions which appear to be different? The scope of such questions is large.

Our reflection involves those general principles which, through logistic choices, have led to the fragmentary or topological systems that currently exist in those fields. The more important systems are considered critically and selected for their essential contribution to the substructure definition or application.

THE NOTION OF SUBSTRUCTURE

Duality of the Notion of Substructure. Substructure is a natural approach to the complex nature of chemical compounds. It constitutes the symbolic expression of an abstract part of the physical object under consideration. Such expression depends upon the chemical representation of the compound and therefore upon the concepts and models designed for its interpretation and description. Explicative or descriptive theories, together with their involvement in fundamental or applied chemistry, are proposed on the basis of global and local properties of compounds. Substructure underlies the local approach of these theories and contributes to the global representation of the compound. Thus, a substructure can be seen as a part extracted from a representation of the whole, and in such cases, the substructural representation is an application, implicit or explicit, of a global perception to a local perception. Alternatively, a substructure can be seen as a prototype identifying a phenomenon according to a given interpretation. Here, the substructure is conceptually perceived as a whole, thus providing representation and interpretation primitives: the extension to a global perception in, for example, molecular mechanics, is an approximation. In fact, reality involves both of these aspects. Duality of the notion of substructure is inherent in its nature: on the one hand, it is regarded as a whole, but on the other hand, it has meaning only within the context from which it was extracted and thus constitutes the abstraction of a local description. Examples of this include use of bond types to indicate bond length or different chemical shifts linked to a single fragment. A substructure can be interpreted either as an intersection within a family of structures, that is, a generalization tool, or as a part of a structure, with all its contextual specificity, that is, a specification tool.

Substructure and Meaning. Chemical theories have evolved by means of an improved knowledge and understanding of the constituent moieties of substances, i.e., the groups, atoms, or particles which make up substances. The representation of compounds by name or other notation has undergone a parallel evolution. Before, it was understood that compounds were made up of constituent elements, the names given to substances often reflected one aspect or another of their physical properties. This was presumably an attempt to point out similarities or dissimilarities between substances. In the absence of analytical methods, such classification sometimes led to the assignment of different names to the same chemical substance, for example, gold discovered in different areas often acquired different names. At the same time, different chemical substances often were assigned the same name. Thus, the term plumbum, for example, has been used for various heavy, low-melting elements, such as tin, lead, and antimony (plumbum candida, plumbum nigrum, and mercasita plumbea, respectively).[5] Without either primitive definitions or constituent analysis, such classifications were doomed to failure.

The discovery of chemical elements allowed for proposals of classification based on constitution. The *Méthode de Nomenclature Chimique*[6] outlined the principles of a dualistic nomenclature for binary combinations of elements in inorganic substances such as oxides or salts[7] and was one of the first approaches toward breaking down of a substance into its component parts. A first notion of a more complex and significant "part" was given by Lavoisier in the term "elément composé" (compound element) which may be considered to be one of the first intuitive definitions of the idea of substructure. The concept of chemical function allows generalization of this approach by identification of behavioral similarities between organic substances, an observation that preceded the discovery of structural similarities. The radical theory, first proposed by Woehler and Liebig, is the basis of the classical fragment approach to the structure of chemical

0095-2338/90/1630-0002$02.50/0 © 1990 American Chemical Society

compounds.

Thus, substructure is linked to a typical behavior or to a specific, chemically significant structural moiety. Additive models propose explanations of properties by establishing specific contributions of different substituents in homologous series. Pattern recognition methods in chemistry correlate the characteristics of substances with their behavior. The part of the structure that is identified by the substructure can then be a set of diverse structural characteristics, with an underlying meaning that is more complex than that possessed by the classical fragments. These structural characteristics may correspond to various degrees of abstraction, but most of the early substructural abstract concepts were designed in the context of substructure search systems. This is mainly due to the fact that the solution to the problem is computer-derived, expressed in terms of graphs, and not linked to an a priori chemical meaning.

Abstract Nature of Substructure. Abstraction is inherent in the notion of substructure. Accordingly, if structure is the abstract representation of a real entity, then substructure is the representation of an abstract entity. Development of a substructure is a process which identifies an aspect of the full structure. Any properties that may be attributed to the substructure can be confirmed or otherwise only in a real structural context. Differences between various substructure systems can be analyzed primarily in terms of the nature and the degree of abstraction.

The complexity of chemical structure analysis and the diversity of parameters and behavior found in large collections of compounds makes it necessary to expand and enlarge the concepts of structural characteristics. Structural chemistry is linked to the paradigm of topological representation of chemical compounds. Graph theory, on the other hand, provides a formal basis for the conceptualization and the expression of different levels of abstraction such as value aggregates or intervals, variable connectivities, or indeterminate sites. Through a generalization process (e.g., the presence or absence of an atom) global structural properties become substructural primitives. In this way, substructure in information systems has evolved toward higher levels of abstraction.

Different types of substructure have been designed for various purposes. In the next section, we propose a classification of substructures according to various criteria and offer a critical analysis. In the last section, we introduce a simple generic substructure model that extends the structural model.

SUBSTRUCTURE AND CLASSIFICATION

A substructure has meaning only in the context in which it is defined. It must be situated relative to a set of substructures or structures and in the framework of one or more fields of investigation.

It is possible to propose and combine different criteria for classification because a substructure can be interpreted from different perspectives. It can belong to several categories or be specific to one.

Classification by Type of Structural Characteristic. This kind of classification is the most frequently used.[8,9] Characteristics may be global or local, the former often being an aggregate of the latter. They can be categorized as follows:

General characteristics include occurrences of specific bonds or atoms, isotopes, charges, number of nodes having a given connectivity, and so on. Topological indices are expressions of different characteristics extracted from the structure graph, for example, sequences of paths of different length.[10]

Functional groups and ring systems were the earliest and are the most classical characteristics. They are a good tool for a first approach to various fields of investigation and represent the basis of systematic nomenclature, additivity models, fragment codes and associated substructure search methods, and structure–activity relationship studies.

Concentric local environments describe an atom and/or a bond environment. There exist various similar definitions which correspond to different semantic approaches. The augmented atom,[11] the hydrogen augmented atom, and the twin augmented atom fragments each describe an atom and one or more of its neighbors. The two latter fragments are more specific than the first because the branching possibilities are limited. The ganglia augmented atom (12) includes bonds connected to focus neighbors. Infra FREL[3] describes an atom and all of its neighbors and, partially or totally, the atom and bond types. The focus connectivity corresponds, therefore, to the exact connectivity of the input structure. FRELs of length 2 exactly describe two layers of atoms around a focus, and generic FRELs are generalized to variable lengths; they may include several types of controlled indeterminants.[4] Environment descriptors are used by the ADAPT system.[13]

Bond environment descriptors, weighted environment descriptors, and augmented environment descriptors are parameters which provide a global description of an environment consisting of two atom layers. These conventional values are associated with smaller fragments which can thus be distinguished, at different levels of specificity, according to their environment. Infra FRELS and FRELs, or generic FRELs, are linked by a similar relationship expressed in terms of topology. Concentric structures were an early approach to the definition of atom groups with or without a priori classical chemical meaning.

Linear substructures describe some kind of information about a path between two atoms. The number of atoms involved and the type of information depend upon the application. The Sheffield group[14] has defined two atoms with varying specifity (simple pair, augmented pair, and bonded pair). Linear sequences[9] include atom or bond sequences and usually comprise between four and six atoms. Connectivity sequences include only the connectivity values of the atoms in the sequence. The "topological torsion"[15] describes more precisely a path with four atoms. Linear substructures such as the topological torsion, which include information about atom attachments (connectivity, neighboring bond values, and so on) can be considered to be hybrid concentric/linear types. In the DARC system, linear substructures are expressed as a subset of generic FRELs.

Other types of linear substructure have a path length that depends upon the structural context. The "Heteropath"[16] describes a path between two non-carbon atoms. Its contribution to SAR studies is interesting. As an example, the pharmacological activity of bisquaternary ammonium compounds is inverted, depending upon the inverse of the number of carbon atoms between the nitrogens.

In the "atom-pair" approach,[17] only the distance between two atoms is retained. In this substructure system, the entire set of atom pairs must be considered in structural analysis, each single characteristic being too generic.

Linear substructures complement concentric substructures, and the conjunction of these two types provides a good tool for substructure searching.[9] Simultaneous use of classical fragments and linear substructures can lead to interesting results. As an example, the deactivating effect of the primary amino group has been invoked to explain discrepancies in the acidities of various compounds.[19]

More generally, a substructure system is composed of different substructure sets, each consisting of a single type of substructure. Different sets may be independent or linked by some relationship. Two independent subsets can correspond to complementary substructural characteristics such as the cyclic and acyclic fragments used in classical screening systems,

or the topological torsion and atom pair which combine specific fixed-length description with generic description for chains of variable length.

Alternatively, two sets of substructures can be expressed hierarchically by specification of a larger environment such as bond- or atom-centered fragments[11] or WRAIR fragments,[36a] by progressive specification of atom and/or bond values starting from the skeleton (Thus, a ring system is described progressively[18] by its skeleton, the heteroatom positions, their values, and, finally, the branching positions.), and by the simultaneous specification of larger environment and undefined position values. (In Eurecas, a FREL inherits the first atom layer from an infra FREL. It specifies infra FREL undetermined positions as well as the second atom layer around the focus.)

Classification by Type of Definition. This type of classification characterizes the kind of operation the designer performs or defines to generate the system substructures.

Manual Definition. Manual definition provides the widest variety of substructures. The most typical manual definition is the query substructure that is developed in most of the substructure search systems. The constraints on this type of substructure are the syntactical constraints of the system language. The descriptive power of the substructural expression is linked to language primitives.

In fragment-based systems, substructures are expressed by logical combination of elementary fragments. In the EXSPEC system,[20] substructures are PROLOG objects defined recursively by means of superatoms or primitives similar to WLN primitives. Graph models allow the highest level of flexibility. The "special substructures" of the first CAS substructure search system[21] and the specific and nonspecific functional groups and radicals of the CIDS system[22] are examples of substructures defined manually by experts for their descriptive power, their use frequency, or an aspect specific to the application. A manual definition step is also used in some QSAR methods, some structure elucidation systems, or generic reaction definitions. Related substructures can represent a priori knowledge of some structure–activity relationship, a structure elucidation constraint, or sets of compounds defined for the purpose of analysis of their biological activities.[23] When manually defined substructures are used, a substructure dictionary is usually created independently from the structural files.

Atom-by-atom searching is the operation of substructure recognition associated with the manually defined substructure.

Substructures Resulting from Operations on Graphs. The first operation to be considered is intersection. One aspect of structural similarity is provided by algorithms which determine the maximal common subgraphs of two or more labeled graphs.[24]

Intersection of a set of structures that have analogous properties can lead to substructures which potentially contribute to these properties. Examples of such properties include biological activity and spectral behavior.[25,26] Intersection also allows one to classify candidate structures in structure elucidation problems[27] and plays an important part in computer-aided synthesis, reaction indexing, identification of reaction sites, and optimization of synthetic pathways.[28]

Structure intersection is managed at either the generic level (the atomic skeleton) or at some specific level, such as atom/bond values, hybridization, number of hydrogen atoms, or chirality. Use is also made of intermediate parameters such as cyclic/acyclic atoms, functional/nonfunctional atoms, atom value aggregates, ring type as atom value, and connectivity.[26,28d] Some systems explicitly distinguish between internal atoms and peripheral atoms. Internal atoms of the common substructure have their parameters specifically defined, while links between the peripheral atoms and the remainder of the structure are defined more loosely. This distinction facilitates the expression of structural changes that characterize a reaction. Such notions of internal and peripheral atoms are inherent to any substructure system.

Another type of operation that provides substructures is progressive construction from basic substructures. This is done during structure elucidation so as to generate all the possible structures consistent with some set of substructural constraints.[29] Substructures ("superatoms",[29a] "components",[29b] or "ELCO"[29f]) implied by spectral data are the building blocks, serving as operands of a union operation on labeled graphs, with or without overlap. Substructure parameters used in intersection are analogous to those used in union. GENOA,[29e] for example, allows specification of single or multiple values for bonds or atoms, intervals of hydrogen atoms, hybridization, configuration, and free valencies.

Union is implicit in substructure search systems. Structures that are retrieved contain the substructures defined in the query, and the appropriate structures are those for which a combination of these substructures leads to a structure that is consistent with the query structure. Overlapping substructures limit potentially incorrect constructions.[3,9]

Definition by Syntactic Types. Substructure search systems have helped to free the concept of substructure from the constraint of an a priori chemical meaning. The computer has permitted the exploitation of classical chemical fragments in large collections of compounds in substructure search as well as pattern recognition systems. Use of computers has shown the limits of this approach and also the advantages of greater abstraction on the basis of the processing of the mathematical entity (the labeled graph) representing the structures. Typical examples of this include procedures such as the Morgan algorithm and the DARC structure generation method, which provide unique structure representations, and hash coding functions based upon labeled graphs.[30] Beginning with the earliest studies on atom-by-atom searching,[31,32] screens focused on any node or edge and systematically extracted from the graph representation of structure have been exploited. Examples of this include bond–atom–bond or atom–bond–atom triplets, with possible further extension to *n*-tuples.

Composition rules allow for definition of different types and subtypes of substructures. The "augmented atoms" proposed by the Sheffield group, the "plain" and "special" bond-centered fragments,[35] and the different types of FRELs used by the DARC system are examples of such definitions. The FREL type, for example, is a generic definition of disjoint classes of FRELs (connectivity 2, 3, 4, ...). They are progressively described through a process involving generalization and specification of atom values or atom connectivity. Infra FRELs constitute an intermediate level. Different classes of substructures (e.g., ELCOs[29f]) can be defined for application purposes.

The specific generic hierarchy of predefined substructure types is similar to general knowledge representations as are found for example in frames[33] or semantic networks.[34]

Two kinds of operations involving prototype-defined substructures can be associated with the recognition operation involved in manually defined substructures. These are extraction and identification. For extraction, for a given type, the procedure accepts an input structure and recognizes instances of the given type. This is a parsing operation. For identification, a comparison of the extracted substructure is made with the members of a collection of previously extracted substructures (databases, test populations, dictionaries, hierarchies, and so on). This can be a simple comparison of descriptors or it can include elucidation of different levels of specificity.

Classification According to Degree of Specificity: Homogeneous and Heterogeneous Substructures. Feldmann and Hodes[36] identify several levels of specificity, including atom value only, bond value only, skeleton and intermediate levels, cyclic and acyclic bonds, saturated and unsaturated bonds, and bond and atom aggregate values. A set of substructures developed by algorithmic extraction is generally homogeneous in that it includes a single degree of specificity[37] to limit the combinatorial possibilities. This type of definition allows easy handling and minimizes storage requirements. However, it limits the potential use of substructure. The substructures developed in this way are not flexible enough to represent the diversity of knowledge as it is found in various applications. For example, in a substructure search, the substructures extracted from a query structure for screening purposes will be at the level of the specificity of the most generic entity within the environment recognized by the substructure, neglecting any specific local information. Consequently, the use of different types of screens is indicated, but this does not always recover the information that was lost. The generic FRELs of the DARC substructure search system are extracted from query FRELs and include both generic and specific aspects. Their elucidation is controlled by infra FRELs which can include different types of sites such as single, multiple, or indeterminant bond or atom values. The ganglia augmented atoms resulting from the specification of bond values of augmented atom peripheral atoms are defined with two levels of specificity for atoms.[12]

Topological torsion (definition of a four-atom chain) and atom pairs (description of a chain of any length) are two complementary sets possessing different degrees of specificity. Various levels of specificity have been proposed for the description of ring systems,[18] and they constitute good tools for retrieval of partial ring systems.

Inclusive and Exact Substructures. An important distinction among substructures can be made by considering their potential for inclusion within a specific structure. Substructures with no branching nodes can be considered to be generic structures, but for all other substructures, their definition contains implicit or explicit specifications for their imbedment within a structure. Definition of a substructure according to topological distance criteria, as in a concentric substructure, for example, permits attachments to peripheral atoms. Examples of the most common explicit parameters which also reflect the level of specificity include free site, free valence, hybridization, number of π electrons, and number of hydrogen atoms. These play an important role in the problem of generation of structural isomers under substructural constraints.[29]

In CHEMICS,[38] for example, this operation is carried out by means of sets of hierarchically defined substructures or components, in which the branching bond or atom type is specified at the lowest level. In the EPIOS system,[29f] combinatorial possibilities are limited owing to the use of overlapping substructures, the ELCOs. Augmented atoms are totally inclusive; the connectivity of the central atom does not express the connectivity of the corresponding atom within the structure. In infra FRELs, on the other hand, the focus describes the exact connectivity, while the generic aspect, which is more restrictive here, is expressed by means of indeterminate positions, as shown in Figure 1. A query FREL may exhibit indeterminate focus connectivity. A search for inclusion is therefore performed by logical ORs between every possible specific connectivity.

GENERIC EXTENSION OF THE GRAPH STRUCTURAL MODEL

The evolution of substructure representation runs parallel to that of historical computer structure representation, which has proceeded from fragment to topological codes. Intermediate representations use symbols for conventional expressions of atoms in specific environments. These can be simple, as in the WLN or the Mechanical Chemical Code,[39] or complex, as in the GREMAS code.[40] The molecular paradigm leads to the synthetic topological representation of various theoretical aspects. Structural chemistry theory and graph theory both permit the deduction of various types of property from a specific representation of the structure or substructure. Some of these global properties, which are common to different specific situations, are in fact explicit primitives for substructures. Examples of such properties include the number of atoms between each two in an atom pair,[17] the number of π-electrons and the connectivity in the topological torsion,[15] and hybridization. The abstraction that is necessary to create substructures results in a generalization of the concepts of structural models, as may be seen in free sites, free valencies, residual valencies, undetermined atoms, or bond values.[4,9,29e,41] These parameters define the limits of the indeterminancy inherent in substructures.

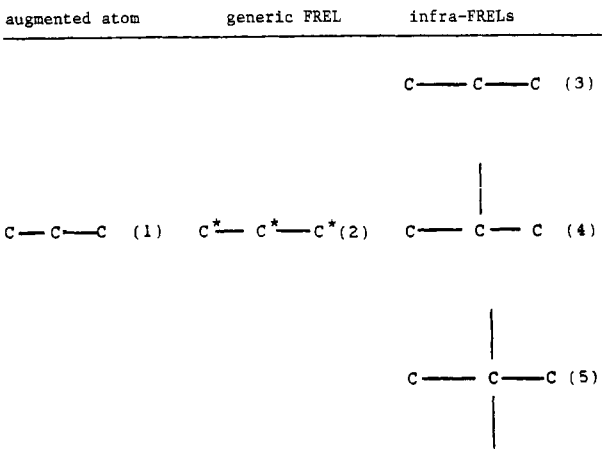

Figure 1. Augmented atom (1) and generic FREL (2) are totally inclusive (stars identify sites with potential attachments; they are implicit in augmented atoms). FRELs 3–5 each elucidate a specific connectivity of the central atom; they are the expression of three disjoint substructural sets.

Substructure system definitions are based on the structural model. We propose that it be extended to a generic model which provides a formal basis for the different concepts that are encountered in this field.

The classical graph model, whether or not it is recognized, underlies all topological approaches. In the DARC system, for example, a structure is expressed by a labeled graph G (X, U, Fx, Fu), where X is the set of nodes, U is the set of edges, Fx is the mapping of X into A, and Fu is the mapping of U into B. A (resp.U) is the set of all possible values of atoms (resp.bonds).

The structural model is still too restrictive to express generic concepts fully. Accordingly, we propose to extend it to a generic model which can recognize a generic substructure by means of a graph G (X, U, Fx, Fu, S) in which X is the set of nodes, U is the set of edges, Fx is the mapping of X into $P^*(A)$, Fu is the mapping of U into $P^*(B)$, and S is the mapping of X into N. $P^*(A)$ (resp.(B)) is the set of parts of A (resp.B) where the empty element has been excluded. N is the set of integers and the function S associates with each node an integer which represents the number of additional attachments allowed on this node.

This constitutes a formal expression for the types of information handled in certain existing systems, such as query structures in the DARC system or DENDRAL substructures. This basic model has been proposed in previous studies[42] to formalize the DARC substructure tools, while unifying their design. The different types of ELCOs used in the EPIOS

structure elucidation system, for example, are subtypes of the generic FREL (cf. Appendix). It can be considered the underlying topological scheme used formally in different areas of the DARC system and can include other types of functions for the description of specific features.

According to this formulation, the structural model appears to be a restriction of the generic model. Elements of A (resp.B) are also elements of $P(A)$ (resp.$p(B)$), and zero (element of N) is the value linked to each node of the structural graph by an implicit S function. The generic model is a unified model for structures, generic structures, and generic substructures. Generic FREL concepts and their application to the Markush representation and to the Markush DARC screening system prototype[44] are all handled by this model. Most of the substructural characteristics described in the preceding sections can be described by this model. Undefined bonds or atoms, bond or atom value aggregates, additional attachment possibilities, and internal and peripheral atoms differentiated by means of the S function are all supported, and it is possible, therefore, to handle a wide range of applications. Other characteristics such as free valence, hybridization, or variable path length between two atoms can be added to the model as required by any particular application.

CONCLUSION

The substructure and its role in chemistry have been described. Substructure represents a typical approach to the complexity of problems of representation. Large-scale representation is possible only through restricted underlying theories. Different types of classification have been proposed, each one emphasizing one particular aspect, depending upon the application. DARC substructures, defined as subtypes of the generic FREL model, have been discussed within their general framework.[2-4] A unified model has been proposed as the basis for further work on substructure representation and classification. This work will be reported in future publications.

ACKNOWLEDGMENT

We express our deepest appreciation for the invaluable help provided by the Editor in correcting this paper.

APPENDIX: ELCO AND FREL CONCEPTS

The ELCO concept and its related, ordered description, defined in the development of the DARC code, deal with an environment which may be specific or generic but is always finite. It is valid for structures and for substructures. In addition, it is a description of the structural information that is captured by the step-by-step propagation of an ELCO and its related order. The first ELCO is concentric and describes fully the environment of the focus of the structure. The subsequent ELCOs however, omit that part of the concentric information that has already been described at a preceding propagation step. The code that is generated is therefore nonredundant, and the FREL is a fragmentation tool which gathers local concentric information together, independent of any larger structural context. The first ELCO, which is rooted in the structure focus, describes the same substructural information as does the FREL which is rooted in the same focus.

Generic or specific substructures used in the DARC EPIOS expert system, which was developed for elucidation of structures from C13 NMR data, have been named ELCOs, rather than FRELs, so as to emphasize their role in the generation of structures through a polyfocalization process.

REFERENCES AND NOTES

(1) Dubois, J.-E.; Laurent, D.; Viellard, H. Systeme de Documentation et d'Automatisation des Recherches de Correlation (DARC). Principaux Generaux. *C. R. Séances Acad. Sci., Ser. C.* **1966**, *263*, 764-767.
(2) Dubois, J.-E. Structural Organic Thinking and Computer Assistance in Synthesis and Correlation. *Isr. J. Chem.* **1975**, *14*, 17.
(3) Attias, R. DARC Substructure Search System: A New Approach to Chemical Information. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102-108.
(4) Dubois, J.-E.; Panaye, A.; Attias, R. DARC System: Notions of Defined and Generic Structures. Filiation and Coding of FREL Substructure Classes. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 74-82.
(5) Crosland, M. P. *Historical Studies in the Language of Chemistry*; Dover Publications: New York, 1978; p 394.
(6) Guyton de Morveau. *Méthode de Nomenclature Chimique* **1987**, op. cit. ref 7.
(7) Lozach, N. *La Nomenclature en Chimie Organique*; *Masson*: Paris, 1967.
(8) Bawden, D. Computerized Chemical Structure Handling Techniques in Structure–Activity Studies and Molecular Property Prediction. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 14-22.
(9) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. The CAS ONLINE Search System. 1. General Design and Selection, Generation, and Use of Search Screens. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93-102.
(10) Randic, M.; Wilkins, C. L. Graph Theoretical Approach to Recognition of Structural Similarity in Molecules. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 31-37.
(11) Adamson, G. W.; Lynch, M. F.; Town, W. G. Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part II. Atom-Centered Fragments. *J. Chem. Soc. C* **1971**, 3702-3706.
(12) Hodes, L. Selection of Molecular Fragment Features for Structure–Activity Studies in Antitumor Screening. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 132-136.
(13) Bruger, W. E.; Stuper, A. J.; Jurs, P. C. Generation of Descriptors from Molecular Structures. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 105-110.
(14) (a) Adamson, G. W.; Bush, J. A.; McLure, A. H. W.; Lynch, M. F. An Evaluation of a Substructure Search Screen System Based on Bond-Centered Fragments. *J. Chem. Doc.* **1974**, *14*, 44-48. (b) Adamson, G. W.; Cowell, J.; Lynch, M. F.; McLure, A. H.; Town, W. G.; Yapp, A. M. Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files. *J. Chem. Doc.* **1973**, *13*, 153-157.
(15) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82-85.
(16) Chu, K. C.; Feldmann, R. J.; Shapiro, M. B.; Hazard, G. F.; Geran, R. I. Pattern Recognition and Structure–Activity Relationships Studies. Computer-Assisted Prediction of Antitumor Activity of Structurally Diverse Drugs in an Experimental Mouse Brain Tumor System. *J. Med. Chem.* **1975**, *18*, 539-545.
(17) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure/Activity Studies. Definitions and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64-73.
(18) Feldmann, R. J.; Heller, S. R. An Application of Interactive Graphics. The Nested Retrieval of Chemical Structures. *J. Chem. Doc.* **1972**, *12*, 48-54.
(19) Klopman, G. Artificial Intelligence Approach to Structure–Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315-7321.
(20) Luinge, M. J.; Kleywegt, G. J.; Van't Klooster, H. A.; van der Maas, J. H. Artificial Intelligence Used for the Interpretation of Spectral Data. Automated Generation of Interpretation Rules for Infrared Spectral Data. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 95-99.
(21) Dayton, D. L. Dynamic Applications of Substructure Searching in a Chemical Information System. Proceedings of the 156th National Meeting of the American Chemical Society: Atlantic City, Sept 1968.
(22) Powers, R. V.; Hill, H. N. Designing CIDS. The U.S. Army Chemical Information and Data System. *J. Chem. Doc.* **1971**, *11*, 30-38.
(23) Nasr, M.; Paull, K. D.; Narayanan, V. L. Computer-Assisted Structure–Activity Correlation. *Adv. Pharmacol. Chemother.* **1984**, *20*, 123-190.
(24) Levi, G. A Note on the Derivation of Maximal Common Subgraphs of Two Directed or Undirected Graphs. *Calculus 9*, 1-12.
(25) Cone, M. M.; Venkataraghavan; McLafferty, F. W. Molecular Structure Comparison Program for the Identification of Maximal Common Substructures. *J. Am. Chem. Soc.* **1977**, *99*, 7668-7671.
(26) Varkony, T. H.; Shiloach, Y.; Smith, D. H. Computer-Assisted Examination of Chemical Compounds for Structural Similarities. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 104-111.
(27) Lipkus, A. H.; Munk, M. F. Automated Classification of Candidate Structures for Computer-Assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 9-18.
(28) (a) Vleduts, G. E. Development of a Combined WLN/CTR Multilevel Approach to the Algorithmical Analysis of a Chemical Reaction in View of their Automatic Indexing. British Library R&D Department Report 5399, London, 1977. (b) Lynch, M. F.; Willett, P. The Production of Machine-Readable Descriptions of Chemical Reactions Using WLN. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 94-96. (c) McGregor, J. J.; Willett, P. Use of Maximal Common Subgraph Algorithm in the Automatic Identification of the Ostensible Bond Changes

Occurring in Chemical Reactions. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 137-140. (d) Bersohn, M. An Algorithm for Finding the Intersection of Molecular Structures. *J. Chem. Soc., Perkin Trans. 1* **1982**, 631-637. (e) Dubois, J.-E. Computer-Assisted Modelling of Reactions and Reactivity. *Pure Appl. Chem. 53*, 1317-1327. (f) Sicouri, G.; Sobel, Y.; Picchiottino, R.; Dubois, J.-E. Système DARC. Localisation des Variations sur l'Invariant d'une Reaction: Concept de Structure Transformante. *C. R. Acad. Sci., Ser. 2*, 523-528.

(29) (a) Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. Applications of Artificial Intelligence to Chemical Inference. 17. An Approach to Computer-Assisted Elucidation of Molecular Structure. *J. Am. Chem. Soc.* **1975**, *97*, 5755-5762. (b) Sasaki, S.-I.; Abe, H.; Hirotie, Y.; Ishida, Y.; Kudo, Y.; Ochiais, S.; Saito, K.; Yamasaki, T. Chemics-E: A Computer Program System for Structure Elucidation of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 211-222. (c) Lipkus, A. H.; Munk, M. E. Combinatorial Problems in Computer-Assisted Structural Interpretation of C13 NMR Spectra. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 38-45. (d) Shelley, C. H.; Hays, T. R.; Roman, R. V.; Munk, M. E. An Approach to Automated Partial Structure Expansion. *Anal. Chim. Acta* **1978**, *103*, 121-132. (e) Carhart, R. E.; Smith, D. H.; Gray, N. H. B.; Nourse, J. G.; Djerassi, C. GENOA: A Computer Program for Structure Elucidation Utilizing Overlapping and Alternative Substructures. *J. Org. Chem.* **1981**, *16*, 1708-1718. (f) Dubois, J.-E.; Carabedian, M.; Dagane, I. Computer-Assisted Elucidation of Structures by Carbonates NMR. The DARC-EPIOS Method: Characterization of Ordered Substructures by Correlating the Chemical Shifts of Their Bonded Carbon Atoms. *Anal. Chim. Acta* **1984**, *158*, 217-233.

(30) Freeland, R. G.; Funk, S. A.; O'Korn, L. J.; Wilson, G. A. The CAS Registry System. II. Augmented Connectivity Molecular Formula. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 94-105.

(31) Ray, L. C.; Kirsh, R. A. Finding Chemical Records by Computer. *Science* **1957**, *126*, 814-818.

(32) Moers, C. N. *Ciphering Chemical Formulas.* Zatopleg System Zator Technical Bulletin 59; The Zator Co.: Boston, 1951.

(33) Minsky, M. A Framework for Representing Knowledge. In *The Psychology of Computer Vision*; Winston, P., Ed.; McGraw-Hill: New York, 1975.

(34) Quillian, M. R. Semantic Memory. In *Semantic Information Processing*; Minsky, M., Ed.; MIT Press: Cambridge, MA, 1968.

(35) Hodes, L. Clustering a Large Number of Compounds. 1. Establishing the Initial Method on an Initial Sample. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 66-71.

(36) (a) Feldmann, A.; Hodes, L. An Efficient Design for Chemical Structure Searching. I. The Screens. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 147-152. (b) Feldmann, A.; Hodes, L. Substructure Search with Queries of Varying Specificity. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 125-128.

(37) Lynch, M. F. Screening Large Chemical Files. In *Chemical Information Systems*; Ash and Hyde, Eds.; Ellis Horwood: Chichester, U.K., 1975.

(38) Abe, H.; Okuyama, T.; Fujiwara, I.; Sasaki, S.-I. A Computer Program for Generation of Constitutionally Isomeric Structural Formula. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 220-229. Funatsu, K.; Migabayashi, N.; Sasaki, S.-I. Further Development of Structure Generation in the Automated Structure Elucidation Program CHEMICS. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 18-28.

(39) Lefkowitz, D. Substructure Search in the MCC System. *J. Chem. Doc.* **1968**, *8*, 166-173.

(40) Rossler, S.; Kolb, A. The GREMAS System, an Integral Part of the IDC System for Chemical Documentation. *J. Chem. Doc.* **1970**, *10*, 128-134.

(41) Dubois, J.-E.; Panaye, A.; Picchiottino, R.; Sicouri, G. Systeme DARC: Structure de l'Invariant d'une Reaction. *C. R. Acad. Sci. Ser. 2* **1985**, *295*, 1081-1086.

(42) Dubois, J.-E.; Attias, R. ITODYS Internal Report, Nov 1979.

(43) Dubois, J.-E.; Sobel, Y. DARC System for Documentation and Artificial Intelligence in Chemistry. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 326-333.

(44) The Markush Search System derived from the substructure search system comprises two kinds of structural screens: specific FRELs (extracted from the invariant part) and generic FRELs describing specific aspects of the invariant structure and generic aspects of the variable structural moieties. Lourdin, C. Traitements des Formules Generiques des Brevets dans le Systeme DARC. Thesis, University of Paris, 1976.

Automatic Processing of Graphics for Image Databases in Science

ROBERTO ROZAS* and HUGO FERNANDEZ

Department of Chemistry, University of Santiago de Chile, Casilla 5659, Santiago 2, Chile

Received April 4, 1989

Generation of a database for automatic characterization, storage, and retrieval of graphic scientific information is presented. The system makes use of a scanner that allows one to digitize any graphic information. The software developed for processing the digitized images generates unique descriptors for representing the images. The system also provides a cubic spline treatment for the stored descriptors to get polynomial coefficients which are used to retrieve any graphic correlation such as a spectrum or a $y = f(x)$ representation. The original image is also compacted for further utilization. The automation provided by this graphic or image database (IDB) makes the classification and retrieval treatment human independent.

INTRODUCTION

Alphanumeric databases offer great benefits in selective retrieval of specific records within huge amounts of information.[1-3] Knowledge communication in science, however, is normally alphanumeric and also graphic and sometimes is essentially graphic. This reality makes attractive the possibility of having a graphic or image database (IDB). The use of IDBs is important in science due to the need to study similar patterns or correlations found by previous researchers or even to establish new correlations. For instance, if we have a graphic pressure-volume representation, a NMR or an IR spectrum, we can retrieve it or retrieve its similar curves according to its shape.

The IDBs known in the literature have been implemented in a way in which the graphic information is characterized by alphanumeric attributes.[4,5] These attributes are externally incorporated to the database by human processing, not deduced by a program, even when several theoretical algorithms for automatic characterization of curves have been proposed.[6]

In this paper we present a system that automatically works out the scientific graphic information necessary for the storage and further selective retrieval of images. This is done with a scanner[7] and a specially designed program for image processing and for graphic representation and management. Representation of the graphic information is based on cubic splines,[8] whose coefficients allow for the comparison of the images in the retrieval process.

GENERAL DESCRIPTION OF THE SYSTEM

Let us suppose we want to put into the IDB a record (imaginary) constituted of alphanumeric and graphic information such as

Electroantennogram correlations for pheromone compounds
J. J. Spencer et al.
J. Volat. Sci. **1980**, *40*, 200-214

0095-2338/90/1630-0007$02.50/0 © 1990 American Chemical Society