the algorithm on the check digit) to see whether the number is a valid one, thereby eliminating most keying errors for registry numbers. Similarly, alphanumeric strings will not be accepted when numerical data are required. After a record is entered at a terminal, a copy of the record as it will be entered into the CHEMFATE file and the record number is displayed on the screen, giving the person entering the data an opportunity to review what has been entered and to make notes on what should be edited. After these corrections are made, a copy of the records are returned to the individual who abstracted the information to recheck the file entry.

The CHEMFATE records have a two-faceted key. The first part is the record number, and the second part is an interrecord line number. CHEMFATE records can be retrieved rapidly by searching ACCT which is keyed by CAS registry number for the desired registry number, obtaining the relevant record numbers and retrieving these records from CHEMFATE.

The report generation program for CHEMFATE has been completed. For a specified CAS registry number this program retrieves and formats all records on the chemical, any of the 22 data types or ID (identification) (see Table IV), or the four groups of data types, namely, chemical dynamic properties, transport properties, laboratory degradation, and field studies and monitoring. Examples of CHEMFATE records appear in Figure 4. It is planned to make both DATALOG and CHEMFATE data bases available to the public.

## ACKNOWLEDGMENT

The support for this work from the U.S. Environmental Protection Agency, Office of Toxic Substances, under Cooperative Agreement CR 806902020 is gratefully acknowledged.

## REFERENCES AND NOTES

(1) Conway, R. A. "Environmental Risk Analysis of Chemicals"; Van Nostrand Reinhold: New York, 1981.

(2) Howard, P. H.; Santodonato, J.; Durkin, P. R. "Syracuse Research Corporations Approach to Chemical Hazard Assessment". In "Environmental Risk Analysis for Chemicals"; Van Nostrand Reinhold: New York, 1981.

(3) Christensen, H. E.; Fairchild, E. J.; Carroll, B. S.; Lewis, R. L., Sr. "Registry of Toxic Effects of Chemical Substances"; NIOSH, U.S. Government Printing Office: Washington, DC, 1980.

(4) Christensen, H. E.; Fairchild, E. J.; Lewis, R. L. "Suspected Carcinogens"; 2nd ed.; NIOSH, U.S. Government Printing Office: Washington, DC, 1976.

(5) Kemp, H. T.; Little, R. L.; Holoman, V. L.; Darby, R. L. "Water Quality Criteria Data Book, Vol. 3 and 5. Effects of Chemicals on Aquatic Life"; U.S. EPA, U.S. Government Printing Office: Washington, DC, 1971, 1973.

(6) Fishbein, L. "Potential Industrial Carcinogens and Mutagens"; U.S. EPA 560/5-77-005.

(7) National Library of Medicine. "Toxicity Data Bank"; U.S. Department of Health, and Human Services: Washington, DC.

(8) Magnuson, V.; Harriss, D.; Maanun, W.; Fulton M. "ISHOW User's Manual"; University of Minnesota: Duluth, MN, 1979.

(9) OHMTADS (Oil and Hazardous Materials Technical Assistance Data System), NIH/EPA Chemical Information System.

(10) For example, WATERDROP (Distribution Register of Organic Pollutants NIH/EPA Chemical Information System) and STORET (STOrage and RETrieval of water quality data), U.S. EPA Office of Water and Hazardous Materials, Washington, DC.

(11) Howard, P. H.; Saxena, J.; Sikka, H. C. "Determining the Fate of Chemicals". *Environ. Sci. Technol.* **1978**, *12*, 398-407.

(12) SRI International. "A Study of Industrial Data on Candidate Chemicals for Testing"; EPA 560/5-77-006; U.S. Nat. Tech. Inform. Serv. PB274264.

(13) Hansch, C.; Leo, A. J. "Substituent Constants for Correlation Analysis in Chemistry and Biology"; New York, 1979.

(14) S. Yalkowsky at Upjohn in Kalamazoo, MI, has collected aqueous solubility data on 2000 nonelectrolyte solutes.

(15) Zwolinski, B. J. and Wilhoit, R. C. "Handbook of Vapor Pressure and Heats of Vaporization of Hydrocarbons and Related Compounds"; Thermodynamics Research Center: College Station, TX, 1971; API44-TRC101.

(16) Perrin, D. D. "Dissociation Constants of Organic Bases in Aqueous Solution"; Buttersworth: London, 1965; IUPAC Chemical Data Series.

(17) Perrin, D. D. "Dissociation Constants of Organic Bases in Aqueous Solution"; Buttersworth: London, 1972; IUPAC Chemical Data Bases: Supplement 1972.

(18) Serjeant, E. P.; Dempsey, B. "Ionisation Constants of Organic Acids in Aqueous Solution"; Pergamon Press: New York, 1979; IUPAC Chemical Data Series.

(19) Hampson, R. F. "Chemical Kinetics and Photochemical Data Sheets for Atmospheric Reactions"; U.S. Department of Transportation: Washington, DC, 1980; Report FAA-EE-80-17.

(20) "Scientific Parameters in Health and the Environment, Retrieval and Estimation: A Requirement Analysis and Examination of Alteratives"; CRC Systems Incorporated: Fairfax, VA, 1981; EPA Contract 68-01-4795.

(21) Lefkovitz, D.; Rispin, A.; Kulp, C.; Hill, H. "EPA Health and Environmental Effects Data Analysis System". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 18-28.

# Fast, Parallel Relaxation Screening for Chemical Patent Data-Base Search

LES KITCHEN and E. V. KRISHNAMURTHY*

Computer Vision Laboratory, Computer Science Center, University of Maryland,
College Park, Maryland 20742

Received September 14, 1981

Described here is an application of the discrete relaxation scheme to search for specific structures and substructures which are included within the generic chemical structure expressions in a chemical patent data base. This scheme can be made highly parallel, since only the local compatibility conditions that are independent are checked, and these checks can be performed simultaneously on a parallel multiprocessor computer, with enormous savings in computation time.

## INTRODUCTION

One of the greatest problems encountered in dealing with the information in chemical patents is the widespread use of generic chemical nomenclature or Markush expressions, where classes of molecules are described which may be either finite or potentially infinite in number, depending upon the constraints placed on the possible position and variety of substituents or other variable characteristics. The economic importance of constructing an efficient computer-based information system for this purpose is now clearly understood.

Most current chemical information systems are widely used for the retrieval of specific structures and of groups of compounds related by their having substructures in common and so are as such inadequate to handle generic structural information. Also, these use essentially node-by-node sequential

* Indian Institute of Science, Bangalore, India.

0095-2338/82/1622-0044$01.25/0

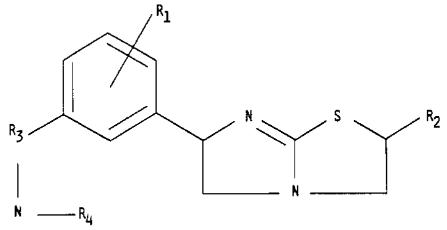


**Figure 1.** Typical generic chemical structure.

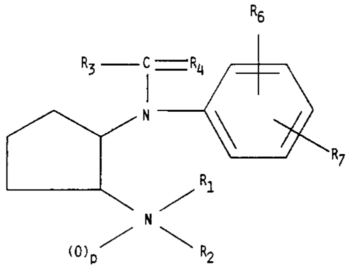


**Figure 2.** Generical chemical structure with conditional statements.

search for matching such queries, using the connection table representation of molecules, with nodes representing the atoms and edges representing the bonds. Accordingly, these techniques are time-consuming.

Our aim in this paper is to provide a brief introduction to a recently proposed method for patent search[1-4] and describe an efficient screening and matching procedure suitable for parallel implementation in a multiprocessor so as to gain enormous speed in the patent examination process.

## MARKUSH STRUCTURES AND PATENT EXAMINATION PROBLEM

In order to make the problem at hand clear, we describe with a few examples the nature of queries in a typical patent examination problem.[1] Two classes of generic structures occur in chemical patents—the delimited class and undelimited class. Examples 1 and 2 below illustrate the delimited class.

**Example 1 (Figure 1).** Here the generic structure is described by introducing certain variable parameters, such as $R_1$, $R_2$, $R_3$, $R_4$, and $R_5$, where $R_1$ = H, Cl, Br, I, $NO_2$, or $CF_3$, $R_2$, $R_4$, $R_5$ = H or unbranched alkyl chain of length 1-4, or $R_4$ and $R_5$ part of a five- or six-membered saturated ring with no other heteroatoms.

$$R_3 = (CH_2)_n \quad (n = 0 \text{ or } 1)$$
$$\text{or } >C{=}O \text{ or } >C{=}S$$

Note that a description, such as the one above, can represent several thousand different but exhaustively enumerable structures.

**Example 2 (Figure 2).** In this example we have (apart from variable parameters) a number of conditional statements on the assignment of variable parameters. In fact, a typical patent contains many more of these conditional statements, which can vary the substituents and also modify the topology (as already described in example 1).[2] Here, $p$ is 0 or 1; $R_4$ is O or S. $R_3$ is 1-3C alkyl, vinyl, 3-6C cycloalkyl, ethoxy, or methoxymethyl. $R_1$ is H or 1-3C alkyl. $R_2$ is $CH_2C_6H_6$ or $(CH_2)_2C_6H_6$ or 3-6C (alicyclic) alkonyl. $R_6$ and $R_7$ are each H, F, Cl, Br, $CF_3$, 1 or 2C alkyl, 1 or 2C alkoxy. Also, when $R_6$ is $CF_3$, $R_7$ is H; when $R_6$ is 1 or 2C alkoxy and $R_7$ is H, the 1 or 2C alkoxy is in the 3rd position; when $R_6$ and $R_7$ are both halogens or 1 or 2C alkoxy, they are present in the 3- and 4- or 3- and 5-positions.

The undelimited class is illustrated by example 3. Here the variable parameters are less well delimited in that generic nomenclature including terms such as alkyl, aryl, etc., are used and are not otherwise qualified. As a result the number of members in this class becomes potentially infinite. This class of problems is very difficult to deal with.[1]

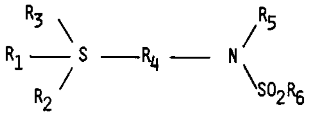


**Figure 3.** Typical undelimited generic formula.

**Example 3 (Figure 3).** Here $R_1$ and $R_2$ are alkoxy or alkyl groups; $R_5$ is hydrogen alkyl or haloalkyl; $R_6$ is alkyl (when $R_5$ is hydrogen), and $R_6$ is alkyl, phenyl, or halophenyl (when $R_5$ is alkyl or haloalkyl); $R_3$ is oxygen or sulfur, and $R_4$ is divalent alkylene or alkyalkylene group.

In a typical patent examination, given the above kinds of generic structures, answers are sought for the following questions.

Q1: Whether a given specific molecule is included in a generic expression.

Q2: Whether certain specified substructures occur within the generic expressions, regardless of whether these lie wholly or only partly within the invariant part of the structure.

Q3: Whether two or more generic expressions have specific molecules in common.

In order to answer these questions we need to design the data base with the following requirements:

(1) A suitable formal language for the description of a generic structure in a patent—This should provide a sufficiently exact and formalized statement of the structural and logical relations in these expressions to permit the algorithmic generation of any member of the delimited class.

(2) As we need to search files with many million chemical compounds, we should have the provision for creating descriptions of localized areas of molecules and a very fast algorithm for search so that node-by-node search is minimized, for finding answers to questions Q1 and Q2 above. Therefore, we need to have a language and a conceptual graph of relations to describe both the invariant and the variable components of the generic chemical structure.

Requirement 1 has been carefully considered in recent papers,[1-4] which discuss the principles and syntax of a generic structure language [GENSAL], which is a formal language, resembling Pascal in many senses. This language facilitates the description of delimited Markush structures by taking as input the connection table representation of the structure diagram and using assignment statements for substituents, other identifiers, position sets for denoting the position of substituent, and conditional expressions which impose conditions on the definitions already made on substituents.

Accordingly, we now have a complete formal description of the delimited Markush structure. This helps us in principle to mechanically enumerate all the required structures contained in a given patent. Therefore, what we need is to devise a very fast procedure to match substructures with the class of generated structures.

In the next section, we describe a procedure called "relaxation" which is very fast and can be implemented on a parallel multiprocessor computer.

## RELAXATION

The relaxation procedure is a parallel iterative method for matching relational structures. This procedure has been used by Kitchen and Rosenfeld[5] for matching structures drawn from several domains. Techniques similar in spirit have also been used by Barrow and Tenenbaum in MSYS,[6] Ullman,[7] and

Waltz.[8] Haralick and Shapiro[9,10] and Mackworth[11] have provided a theoretical analysis of such techniques, while Gaschnig[12] and McGregor[13] have made experimental comparisons among various methods. The structure matching problem can be described as follows: Suppose we have a structure (called W, the "world") described in terms of its parts, their properties, and the relations between them. In the current application, W would be a chemical structural formula, the parts being atoms, their properties being their elemental identities (and possibly other features such as oxidation state), and the relations would be the various sorts of chemical bonds between the atoms. These bonds could be single, double, or triple bonds between pairs of atoms or higher order bonding among groups of atoms as occurs, for example, in a benzene ring. In W we are searching for instances of a given substructure (called M, the "model") which is described in the same terms.

To apply the relaxation procedure, we build an actual or simulated network of parallel processors based on the structure W. That is, we assign one processor to each node (atom) in the structure and make interprocessor connections which correspond to the relations (chemical bonds). Each processor is given a complete description of the sought substructure M, although it can communicate only with those processors to which it is directly connected.

During the processing each processor maintains a list of *labels*, the labels referring to the atoms of M with which this atom of W might possibly be identified. This list is initially set to contain those matches which are possible considering only intrinsic properties of each atom. The relaxation process is itself iterative. On each iteration, every processor eliminates from its list those labels for which there is no possible consistent labeling of neighboring nodes. The iterations continue until no further eliminations can be made.

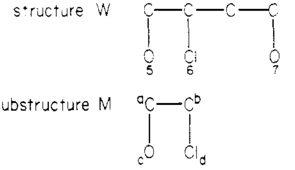
To illustrate this process, we use a very simple example, since any substantial matching problem is far too cumbersome to be gone through in detail by hand.

**Example 4.**

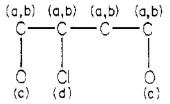
structure W $CH_2OH$–$CHCl$–$CH_2$–$CH_2OH$

substructure M >COH—CCl<

Ignoring hydrogen atoms, and unspecified bonds, and numbering the other atoms for reference, we have

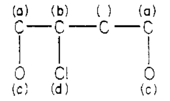


The initial labeling of W with possible matchings gives

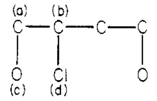


since only carbon can be matched with carbon, oxygen with oxygen, and chlorine with chlorine.

In order to bear the label a, an atom must be bonded to an atom which can bear the label b and to an atom which can bear the label c. Thus, on the first iteration, we can eliminate the label a from carbons 2 and 3. However, carbon 4 retains this label. Similarly, label b can be eliminated from carbons 1, 3, and 4. No other eliminations are possible on this round. So after the first iteration we have



On the second iteration, we can eliminate the label a from carbon 4, since it has lost support from carbon 3. On the third and last iteration, oxygen 7 loses the label c, because carbon 4 has now lost label a. No further eliminations are possible, leaving the structure W correctly labeled with the names of the matching atoms of M.



Because the computations at each node can proceed simultaneously, this relaxation process is eminently suited to implementation on a properly configured multiprocessor computer, permitting many-fold reduction in computation time. Even if the relaxation process is run on an ordinary sequential computer, it still has the advantage that it does not suffer from the "thrashing" behavior which is a problem with many matching techniques based on backtracking.[8,12] Notice that while the computations at each node are purely local, during successive iterations of the process, information is able to propagate through the structure. In the example, the loss of label b on carbon 3 in the first iteration causes the loss of label c on oxygen 7 two iterations later.

One should be aware, however, that relaxation is, from a theoretical point of view, merely a screening process. There are certain circumstances under which the results of the relaxation process indicate that a match may be possible when in fact no match exists. To take an extreme example, consider the matching of the substructure cyclopropane to the structure cyclobutane. It is obvious that no match exists, but since the molecules are identical at a local level, the relaxation procedure is unable to make any eliminations.

However, relaxation is a *safe* screening procedure in that its failures will all be of the above type, of suggesting a match which cannot exist. It is impossible for it to fail in the other way by rejecting a match which does exist. Stronger conclusions than this can be made. The relaxation procedure can stabilize in labelings of only three types: Firstly, we have the situation in which all atoms in W lose all labels. In this case we can safely conclude that no match exists. Secondly, if every node in W has at most one label and every label from M is used just once, then that labeling describes the unique embedding of M into W. Third are results with some remaining ambiguity, in that some atom in W may bear more than one label or some labels from M may be used more than once. Such situations can arise either when there are multiple embeddings of M into W (from actual multiple instances or from internal symmetries of M) or when no match exists, but its refutation requires the use of global evidence, as in the cyclopropane–cyclobutane example above. Such ambiguity cannot be resolved by the relaxation technique, and other methods must be used. One way is to select a single ambiguous labeling and split the matching problem into several subproblems, identical except that in each the ambiguous labeling is resolved in a particular way. If the original problem had a solution, then it must exist in one of the subproblems, and the relaxation process can be applied to them, with recursive use of the splitting technique if any of the subproblems remain ambiguous after the relaxation. This is essentially a combination of relaxation with backtracking, and a similar technique has been used successfully by Barrow and Tenenbaum[6] on other matching problems. Other than this, one could use any standard matching technique, restricting its search to those labelings remaining after the relaxation has finished.

Let us emphasize that while relaxation is a screening procedure, our experience indicates that it is often sufficiently effective to be used alone. That is, the result belongs to one of the first two types described above—either the matching

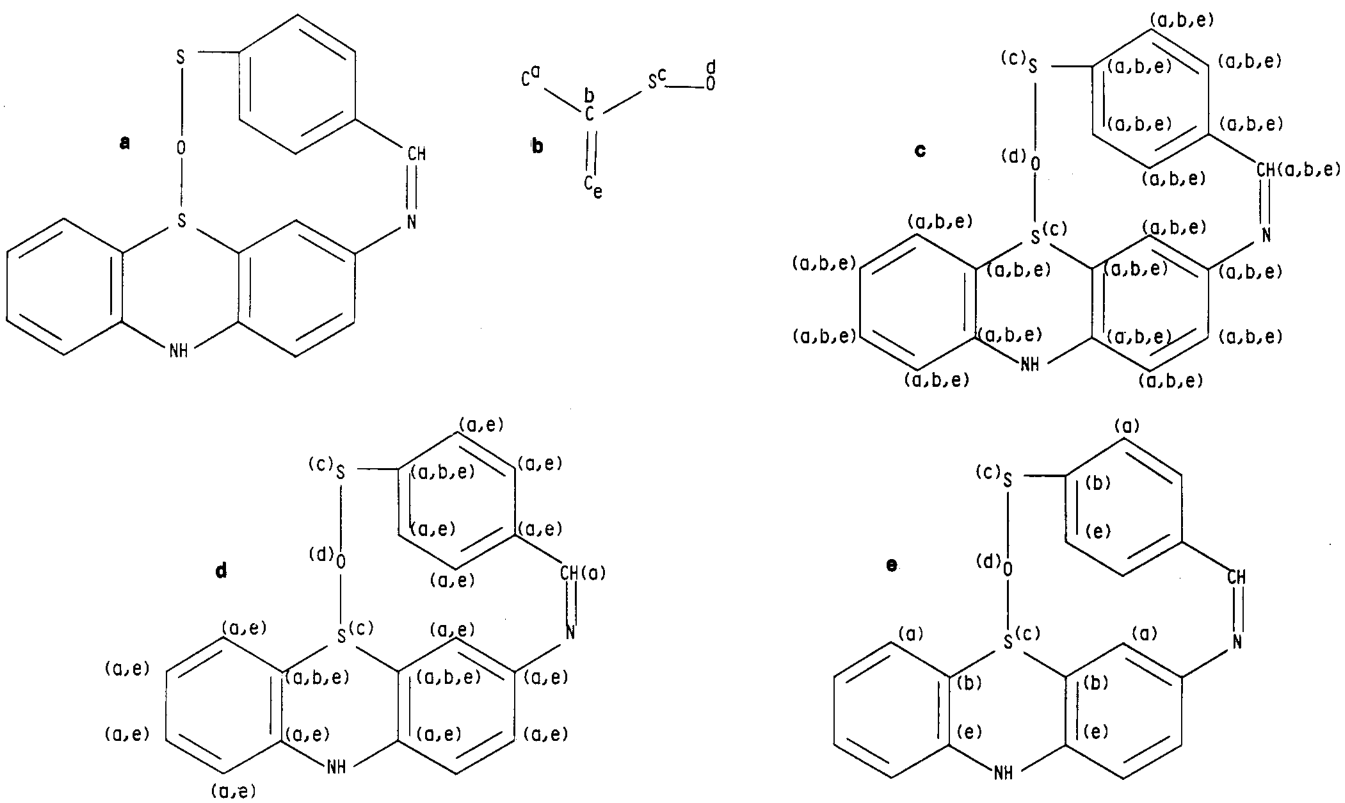


Figure 4. (a) World structure for relaxation matching; (b) model, with atoms arbitrarily labeled; (c) world structure initially labeled; (d) world structure after first iteration of relaxation; (e) world structure after second iteration of relaxation.

is rejected or a unique match is found. In cases where ambiguity remains, we have always found it to be very slight; almost all labelings are rejected by the relaxation process, leaving little work to be done by any subsequent process of disambiguation.

We now give a more substantial example of the relaxation process.

**Example 5.** The World W and Model M for this example are shown in Figure 4a,b, the results after the initial labeling in Figure 4c, and the results after the first and second iterations in Figure 4d,e. Notice that this final labeling is the union of the three overlapping instances of M in W so that a subsequent case analysis would be needed to separate these three instances. However, this analysis would have very little work to do.

The results above were produced by a computer program, written by one of the authors (L.K.) in the programming language Pascal, which implements the relaxation procedure. The program accepts a relational description of a world structure in a symbolic notation. It then accepts descriptions of several model structures in the same notation and attempts to match each one with the world structure. Potentially, the program can handle matching problems with hundreds of nodes and relations of arbitrarily high order, although we have used it only for problems with a few score nodes and relations of order up to six.

The relaxation procedure as it stands at present can only be used for matching a specific substructure against a specific structure. It therefore cannot be used directly for matching generic chemical structures. However, as mentioned above, the GENSAL notation lends itself to the explicit enumeration of the specific structures implicit in a given generic formula, and these specific structures can be matched by using the current relaxation technique. The potential speed of relaxation is important in this approach, given the large number of specific structures that can be generated from a given generic formula. A better approach would be to modify the relaxation procedure so that it can be used for the direct matching of generic formulas. This would require processing considerably more sophisticated than that done at present but should be quite feasible.

## CONCLUSIONS

We have introduced the problems of chemical and patent data-base search and have presented a general purpose technique for structure matching problems (namely, relaxation) that can be applied to chemical patent search. Relaxation can be implemented on a suitably configured multiprocessor computer, with great gains in computational speed over more conventional approaches. Advances in computer technology have made the construction of such machines quite feasible. ZMOB,[14-16] for example, is a network of 256 microcomputers which permits arbitrary intercommunication between them. It is currently in the early stages of construction.

The results presented here should be considered as preliminary and exploratory in nature. For the design of a full-fledged chemical patent information system which is to be implemented on a multiprocessor, we have to consider a number of other aspects. However, it is clear at this point that the approach for this problem should be to design (1) a suitable formal language for patent description and (2) a fast parallel scheme for searching and matching substructures with parts of the generic structure. In this paper, we have considered only the latter aspect. While the relaxation process as it stands can be used for chemical structure matching, it should be possible to modify the technique to make it even more useful. Currently the relaxation process is based on a simple, but quite general relational framework. Considerable savings in time and storage could be realized by using a relational framework more specialized to the chemical domain. For example, chemical bonds are almost always symmetric relations; higher order relations describing ring formation and fusion entail other, lower order relations. Development along these lines should also make possible the extension of this technique to the direct matching of generic structures, without the explicit enumeration of particular structures. Such extensions to the relaxation procedure are important subjects for future research.

## ACKNOWLEDGMENT

The support of the National Science Foundation (Grant MCS-79-23422) is gratefully acknowledged, as is the help of Janet Salzman in preparing this support.

## REFERENCES AND NOTES

(1) E. V. Krishnamurthy and M. F. Lynch, "Analysis and Coding of Generic Chemical Formulae in Chemical Patents", *J. Inf. Sci.*, in press.

(2) M. F. Lynch, J. M. Barnard, and S. M. Welford, "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 1. Introduction and General Strategy", *J. Chem. Inf. Comput. Sci.*, **21**, 148 (1981).

(3) J. M. Barnard, M. F. Lynch, and S. M. Welford, "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 2. GENSAL: A Formal Language for the Description of Generic Chemical Structures" *J. Chem. Inf. Comput. Sci.*, **21**, 161 (1981).

(4) S. M. Welford, M. F. Lynch, and J. M. Barnard, "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 3. Chemical Grammars and Their Role in the Manipulation of Chemical Structures", *J. Chem. Inf. Comput. Sci.*, **21**, 161 (1981).

(5) L. Kitchen and A. Rosenfeld, "Discrete Relaxation for Matching Relational Structures", *IEEE Trans. Syst. Manage. Cybern.*, **SMC-9**, 869-874 (1979).

(6) H. G. Barrow and J. M. Tenenbaum, "MSYS: A System for Reasoning about Scenes", SRI AI Center, Menlo Park, CA, 1976, Tech. Note 121.

(7) J. R. Ullman, "An Algorithm for Subgraph Isomorphism", *J. Assoc. Comput. Mach.*, **23**, 31-42 (1976).

(8) D. Waltz, "Understanding Line Drawings of Scenes with Shadows" in "The Psychology of Computer Vision", P. H. Winston, Ed., McGraw-Hill, New York, 1975, 19-92.

(9) R. M. Haralick and L. G. Shapiro, "The Consistent Labelling Problem: Part I", *IEEE Trans. Pat. Anal. Mach. Intel.*, **PAMI-1**, 173-184 (1979).

(10) R. M. Haralick and L. G. Shapiro, "The Consistent Labelling Problem: Part II", *IEEE Trans. Pat. Anal. Mach. Intel.*, **PAMI-2**, 193-203 (1980).

(11) A. K. Mackworth, "Consistency in Networks of Relations", *Artif. Intel.*, **8**, 99-118 (1977).

(12) J. Gaschnig, "Experimental Case Studies: Backtrack vs Waltz-Type vs New Algorithms for Satisficing Assignment Problems", in Proceedings of the Second National Conference of the Canadian Society for Computational Studies of Intelligence, University of Toronto, Toronto, Ontario, July 1978, pp 268-277.

(13) J. J. McGregor, "Relational Consistency Algorithms and Their Application in Finding Subgraph and Graph Isomorphisms", *Inf. Sci.*, **19**, 229-250 (1979).

(14) C. Rieger, J. Bane, and R. Trigg, "ZMOB: A Highly Parallel Multiprocessor", University of Maryland, College Park, MD, May 1980, Computer Science Tech. Report 911.

(15) C. Rieger, R. Trigg, and J. Bane, "ZMOB: A New Computing Engine for AI", University of Maryland, College Park, MD, March 1981, Computer Science Tech. Report 1028.

(16) C. Rieger, "ZMOB: Hardware from a User's Viewpoint", University of Maryland, College Park, MD, April 1981, Computer Science Tech. Report 1042.

# Structure Evaluation Using Predicted $^{13}$C Spectra[1]

CHRISTOPHER W. CRANDELL, NEIL A. B. GRAY, and DENNIS H. SMITH*

Department of Chemistry, Stanford University, Stanford, California 94305

Received August 14, 1981

A computer program is described for predicting $^{13}$C NMR spectra of organic compounds and for determining the similarity of the predicted spectra to an observed spectrum. The program utilizes a data base containing representations of the stereochemical substructural environments of resonating nuclei, together with their chemical shifts. Given the observed spectrum of an unknown compound and a set of structural candidates for the unknown, the predicted spectra are matched with the observed spectrum, and a score reflecting the degree of matching is calculated. Alternative methods for matching spectra and computing scores are discussed and evaluated using several examples. A matching and scoring function which takes into account the limitations of the data base has proven to yield the best performance.

## INTRODUCTION

In recent years, a number of computer programs have been developed that can construct hypothetical candidate structures for an unknown molecule.[2-6] Gribov[7] and Hippe[8] have recently reviewed such programs and related computer systems. Typically, these programs work by taking a set of substructural fragments, either identified by the chemist or automatically inferred from spectral data and assembling these in all possible ways. The chemist can thus be provided with a set of candidate structures each compatible with all of the more readily interpretable chemical and spectral data. Most structure generation programs work solely in terms of constitutional (topological) isomers; recently, however, algorithms for the constrained generation of configurational stereoisomers have been perfected, thus allowing for more complete structure elucidation.[9]

Such a set of candidate structures, as constitutional or stereoisomers, can in itself be of value to the chemist. Examination of the possible structures can help identify additional spectral or chemical experiments that would resolve remaining alternatives. However, it is also possible for the computer systems to assist in the process of evaluation of the candidates. Typically, programs for candidate structure evaluation are concerned with predicting *spectral* properties for candidate structures and comparing predicted and observed spectra. Differences between predicted and observed spectral properties can be used directly to eliminate candidate structures, or more conservatively, such differences can be captured in some measure of spectral (dis)similarity that is then used for rank ordering the candidates.

Spectral prediction and evaluation algorithms must be capable of processing large numbers of structures of closely related form. This necessarily constrains the type of approach that can reasonably be adopted. In principle, it is possible to use ab initio or semiempirical quantum mechanical methods to compute certain spectral properties of candidate structures, and indeed some limited experiments have been made using MINDO-level semiempirical quantum mechanical methods

0095-2338/82/1622-0048$01.25/0 © 1982 American Chemical Society