Sophisticated Algorithm for Automatic Extraction and Analysis of Substituent-Induced Chemical Shift Differences on ^{13}C-NMR Spectra

Lingran Chen[†] and Wolfgang Robien[*]

Department of Organic Chemistry, University of Vienna, Währinger Strasse 38, A-1090 Vienna, Austria

Received December 23, 1992

A novel approach for quantitative studies of the substituent effects on chemical shift values of ^{13}C-NMR spectra has been developed. The algorithm can automatically extract and analyze the information of substituent-induced chemical shift differences of carbon atoms from all the possible structure pairs in a large database. Substructural fragments can also be selected to indicate which carbon atoms in the structure pairs should be taken into account. The program is a useful tool for a systematic investigation of the influences of substituents on ^{13}C-NMR chemical shift values.

INTRODUCTION

One significant achievement of computer applications in chemistry is the establishment of various spectral database systems, which are now routinely used by spectroscopists and chemists. However, the retrieval commands are mainly designed to supply the user with the information as given in the literature as an answer to the query question. Only a few techniques, such as database-oriented approaches for predicting ^{13}C-NMR spectra,[1–3] are known to be able to extract information from the reference material for an unknown compound. On the other hand, many frequently-encountered problems in chemistry remain because the data available in the database cannot directly be used as solutions, and the manual handling of those problems would be too time-consuming. For example, it is well-known that substituents have strong influences on chemical shifts of ^{13}C-NMR spectra of organic compounds. However, no systematic evaluation of substituent-induced chemical shift differences (SCSD) based on a large reference data collection has been reported so far. The current contribution is a new algorithm which can automatically extract and analyze the SCSD values of ^{13}C-NMR spectra from a large database in order to give answers to various questions concerning substituent effects on ^{13}C-NMR resonance lines. In this paper the algorithm is described in detail, and some examples are given.

BASIC PRINCIPLE

The presented SCSD approach for extraction and analysis of SCSD values can be regarded as a significant extension of our new methodology for automatic calculation of chemical shift differences from two given structures and their ^{13}C-NMR spectra.[4] The main difference of the two methods is that the latter deals with only two well-defined structures, while the SCSD approach does not need to know explicitly the exact structure representations; only the two different substituents (S_1 and S_2) under investigation must be given. The program automatically determines all the possible structure pairs (S_1–R_1, S_2–R_2) from the database, where R_1 and R_2 can be any type of partial structures with at least one carbon atom and they must be isomorphic to each other, i.e., $R_1 = R_2$. As soon as such structure pairs are obtained, the corresponding SCSD values can be easily calculated. Finally, the obtained SCSD-data are analyzed, offering the information about the effects of substituents on chemical shift values.

The main procedures of the algorithm can be described as follows:

(1) Accept the structures of the substituents (S_1 and S_2) and optionally a substructural unit from the user.

(2) Create a table containing the molecular formulas and their corresponding record addresses from the database(s) used.

(3) Rearrange molecular formula table according to the number of carbon and hydrogen atoms in each molecular formula.

(4) Determine all the possible structure pairs according to the molecular formulas as given by S_1 and S_2 and their difference.

(5) Perform screening utilizing three-atom and ring-size information generated from substituents and the fragment under investigation and delete incompatible structure pairs.

(6) Deal with each structure pair.
- (a) Read in the structural and spectral information from the database.
- (b) Perceive substituents 1 and 2 in structures 1 and 2, respectively.
- (c) Perceive the fragment, if given, in structure 1.
- (d) Check for isomorphism of R_1 and R_2.
- (e) Calculate SCSD values for each carbon atom to be considered.

(7) Collect and analyze obtained SCSD values.

The following sections deal with a detailed discussion of the most decisive items of the program operation, which are necessary to get the desired results within a reasonable time. The algorithm must be able to extract the desired SCSD values from a large reference data collection according to very restricted information. The most severe limitation is the large number of structure comparisons which are sometimes necessary because of the fuzzy nature of the request.

INPUT MODES

The SCSD program has four different input modes that allow it to deal with different types of problems. A pseudoatom

[†] On leave from the University of Science and Technology of China, Hefei, Anhui 230026, The People's Republic of China.

Table I. Statistical Information on Chemical Shift Difference Values Induced by a Cyclopropyl-Substituent (Databases Used, A B C F H J K L M; Computing Time, 3 h on a Silicon Graphics 4D25-Workstation)

all possible structure pairs found = 791 823
total no. of structure pairs used = 546

	differences of chemical shifts in each position							
		entries used[b]	average values					entries used[b]
positions considered[a]	highest values	1–2	plus	hits	minus	hits	lowest values	1–2
0	26.7	12901B–5795C	15.3	483	–3.7	1	–3.7	6425B–5227B
1	14.4	14504A–26176B	3.9	307	–2.9	228	–11.8	1486A–5460A
2	3.2	25946A–5154A	0.6	148	–1.1	477	–4.8	8306A–6253A
3	3.5	1280A–1283A	0.7	127	–1.3	371	–4.9	24974B–6060A
4	2.1	8894B–5274C	0.5	81	–0.5	172	–3.4	14504A–26176B
5	0.5	1280A–1283A	0.3	9	–0.2	39	–1.0	23262B–8647A
6	0.3	23422B–7046B	0.2	3	–0.2	13	–0.7	2146B–1093F
7	0.1	25547B–11317B	0.1	2	0.0	2	–0.1	1283A–130B

[a] "Positions considered" indicates the distances between the atom holding the substituent and the other carbon atoms. [b] The two "entries used" columns show the structure pairs (record address and database identifier) corresponding to "highest values" and "lowest values", respectively.

is used to indicate the linking position in substituents and the substructural unit: (1) input of one substituent, (2) input of two substituents, (3) input of one substituent and one substructural unit, and (4) input of two substituents and one substructural unit.

CREATING THE MOLECULAR FORMULA TABLE AND SELECTION OF POSSIBLE STRUCTURE PAIRS

In order to avoid time-consuming duplicate operations on database files, a table containing all the molecular formulas available in the database is first built up. The molecular formulas in the table are arranged according to the number of carbons and hydrogens. Now all possible structure pairs are generated from this molecular formula table using the difference of the molecular formulas derived from the substituents entered and their elemental composition. The number of the possible structure pairs depends upon the size of the reference data collection used and the substituents under investigation. For some substituents, more than 700 000 pairs of structures were generated using our present databases containing about 81 000 ^{13}C-NMR reference spectra.

SCREENING OF THE DATABASES

The handling of all possible structure pairs is the most time-consuming process during the whole procedure, because many operations on database files and very frequent atom-by-atom matching are involved. Therefore, deleting unreasonable structure pairs before the above process is invoked can dramatically reduce the CPU time. In the current implementation, our strategy is based on three-atom screens and ring-size screens. There are two different situations.

(a) No fragment is given: If at least one of the two substituents contains three or more non-hydrogen atoms (except pseudoatom), the above strategy can be used. The three-atom screens and ring-size screens are generated from the given substituents and used to perform the screening in the databases. Those structure pairs which are incompatible with the obtained screening results are then deleted from the structure pair list.

(b) The fragment is given: In this case, instead of calculating screens directly from the input substituents and fragment, respectively, the fragment is first combined with each entered substituent to form two larger structures. Screens are then calculated from these new structures. The following process is the same as described in a.

By using this strategy, the structure pair list can be dramatically reduced and the efficiency of the algorithm is greatly enhanced, because of less atom-by-atom matching steps.

PERCEPTION OF SUBSTITUENTS AND SUBSTRUCTURAL UNIT

Perception of substituents in structures plays an important role in the whole procedure. There are two functionalities of the MCSS algorithm[5] used in this program. First, MCSS-technique is used to check both structures within a structure pair for the occurrence of substituents 1 and 2, respectively. The lack of the corresponding substituent in a structure pair allows further processing of the current pair to be terminated immediately. Second, if both structures in a pair possess substituents 1 and 2, respectively, all of the mappings of substituents upon their corresponding structures are detected and stored for further processing. Using a substructural unit as an additional condition, the MCSS algorithm is used to detect it only within structure 1 of the current pair. The presence of this fragment within structure 2 can be determined by checking the condition $R_1 = R_2$ during the next step. The successful mappings of the fragment upon structure 1 will be used in comparison of structures 1 and 2 and in organizing the output of the result.

COMPARISON OF TWO STRUCTURES IN EACH STRUCTURE PAIR

Each compound in a structure pair may possess more than one substituent under investigation; therefore, it is usually not known which part corresponds to the substituent under consideration and which part corresponds to R, because R may also contain the mapping(s) of the considered substituent. It is apparent that an atom-by-atom matching algorithm cannot be used to compare directly two structures within a possible structure pair. These are the main difficulties in checking the condition $R_1 = R_2$. The key to the solution of this problem is to select correctly the starting points for the comparison of two partner structures. The atoms in structures 1 and 2 which match to the pseudoatoms in substituents 1 and 2, respectively, are selected as starting pairs for comparison of the two structures. In the case of considering a fragment,

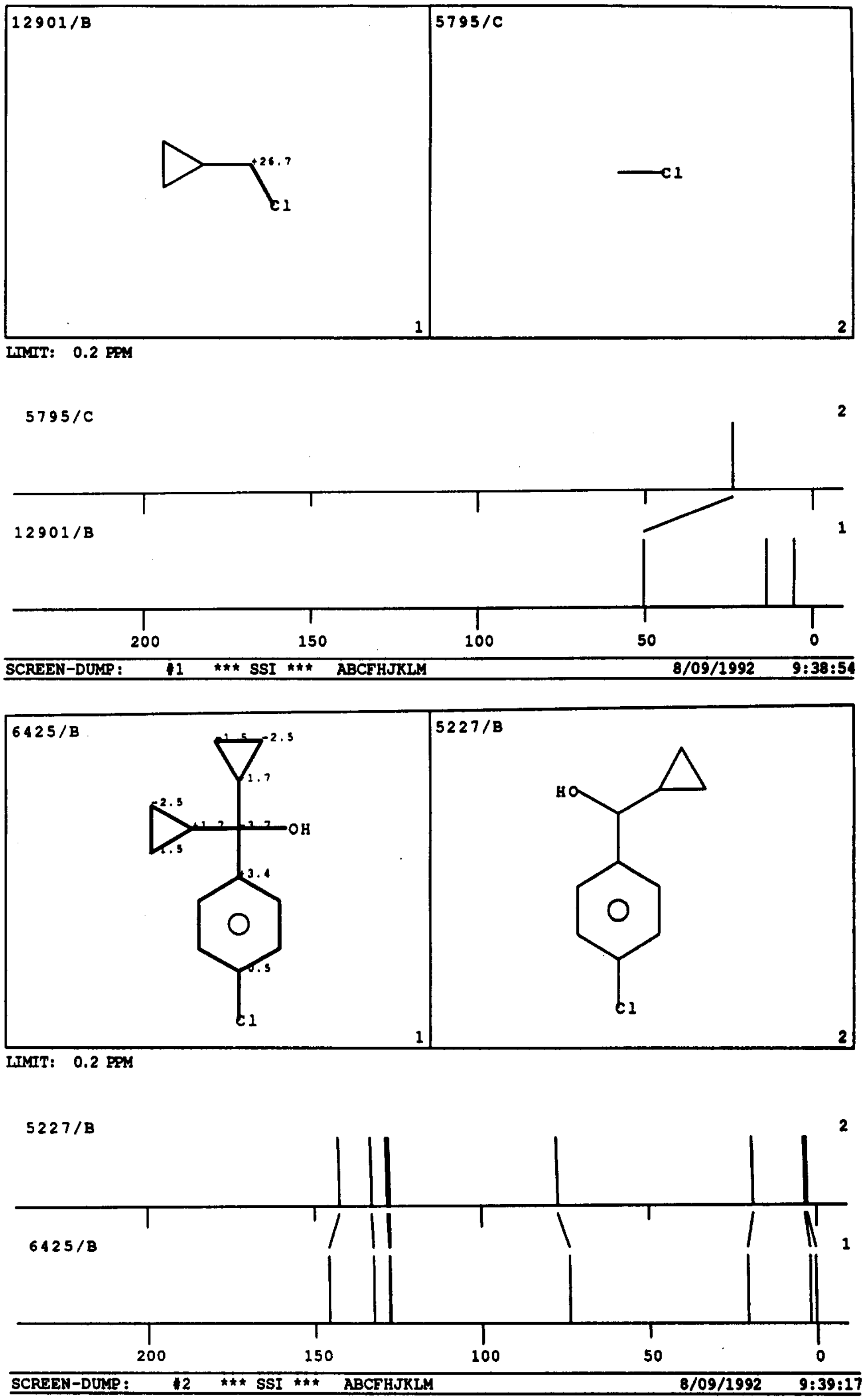

Figure 1. Two structure pairs contributing to the largest SCSD range. The largest positive SCSD value (26.7 ppm) is found for structure pair (12901B–5795C) (top). The SCSD values are inserted into the structural diagram.

only those atom candidates in structure 1 which have led to the successful matching between the fragment and structure 1 will be used as starting points. In order to check the condition $R_1 = R_2$, one possible S_1 part is removed from structure 1, leading to the R_1 part. Structure 2 is manipulated in the same way to generate R_2. Then, the MCSS algorithm is used to check the isomorphism of R_1 and R_2. The successful matching of them leads to the calculation of SCSD values of the desired carbon atoms; otherwise, a new starting pair is tried. This process is repeated until the condition $R_1 = R_2$ is met or all the possible starting pairs have been handled.

OUTPUT OF THE RESULTS

The screen output contains a table showing the statistical information of the obtained SCSD data. For input-mode 1 or 2, the result is arranged according to the distances between the atom holding the substituent investigated and the other carbons in increasing order. This method allows one to collect

Table II. Statistical Information on Chemical Shift Difference Values Induced by a Hydroxy-Substituent (9,10-Anthraquinone Skeleton Structures Retrieved First from Databases A B C F H J K L M; Computing Time, 50 s on a Silicon Graphics 4D25-Workstation)

all possible structure pairs found = 70
total no. of structure pairs used = 11

	differences of chemical shifts in each position							
		entries used[b]	average values					entries used,[b]
pisitions considered[a]	highest values	1–2	plus	hits	minus	hits	lowest values	1–2
1	39.0	22474B–1682A	36.2	11			34.2	1686A–2428A
2	–7.1	22474B–2428A			–8.9	11	–10.3	1686A–1682A
3	5.9	25007A–23196A	4.0	11			2.3	1678A–1682A
4	–3.4	25007A–23196A			–6.2	11	–8.0	1678A–2428A
5	0.2	1678A–1682A	0.2	2	–1.7	9	–3.3	22474B–2428A
6	–0.3	29507B–1682A			–1.3	11	–2.3	25007A–23196A
7	0.6	1686A–1682A	0.5	6	–0.3	5	–0.8	1685A–2428A
8	0.7	22474B–1682A	0.3	6	–0.4	5	–1.0	1686A–2428A
9	0.9	1685A–2428A	0.5	10	–0.2	1	–0.2	25007A–23196A
10	0.6	1686A–2428A	0.4	8	–0.1	3	–0.3	22474B–1682A
11	0.4	22474B–1682A	0.2	3	–0.5	8	–1.3	1678A–2428A
12	0.5	22474B–1682A	0.3	2	–0.3	9	–0.8	1685A–2428A
13	5.4	22474B–1682A	4.6	11			3.5	29507B–2428A
14	–16.4	1686A–1682A			–18.4	11	–20.4	22474B–2428A

[a] "Positions considered" in this table correspond to the carbon numbering as given in Figure 2. [b] Cf. footnote 2 of Table I.

all the possible structure pairs which are consistent with the substituent(s) under consideration. As an example, we investigate the general behavior of a cyclopropyl group as a substituent in all the possible structure pairs available in the database. By using input mode 1, we obtained the result as shown in Table I.

Table I gives the information about the highest, average, and the lowest SCSD values. Each highest and lowest SCSD values are accompanied with their corresponding structure pair information, allowing easy access to the reference data. The average SCSD values are output in two different forms according to the user's choice; one is the total average SCSD values,[6] and the other is obtained by calculating average SCSD values for positive and negative SCSD values separately, as shown as Table I. In both cases, the average SCSD values are given together with the total number of hits used in the calculation. There are 546 structure pairs, which contribute to the final result in this example.

The SCSD values of the α-position have a range of over 30 ppm, from +26.7 ppm for structure pair 12901/B-5795/C (cf. Figure 1, top) to –3.7 ppm for structure pair 6425/B-5227/B (cf. Figure 1, bottom). From Table I it can be easily seen that the influence of a cyclopropyl substituent on the chemical shift values decreases with the increase of the distances from the substituted position, as expected.

SCSD values appearing at carbon atoms with the same distance from the linking position usually vary within a considerably large range, as can be seen from Table I. This fact indicates that the influence of a substituent upon the chemical shift values depends also heavily on the structural environment of the carbon atom considered. Therefore, in many cases, the user may prefer to investigate some specific carbon atoms in a certain fragment, e.g. a benzene ring system; this can be easily done by using input mode 3 or 4. The program uses this fragment information to select only those structure pairs which contain both substituents and the fragment under investigation. Furthermore, only those SCSD values of the carbon atoms within the substructural unit are calculated, and the result is arranged according to the positions of the carbons within the query fragment.

For example, in order to investigate the effects of a hydroxy-group on the carbon chemical shift values in the 9,10-anthraquinone fragment, we use input mode 3: Input of a hydroxy group as substituent and 9,10-anthraquinone (cf. Figure 2) as fragment. The result is shown in Table II.

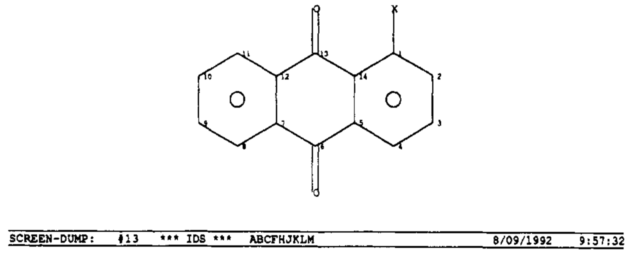

Figure 2. 9,10-Anthraquinone skeleten as fragment used in example 2. X is a pseudoatom indicating the position where the hydroxy substituent is located.

In this example, 11 structure pairs contribute to the final result. The SCSD ranges of all the carbon atoms investigated are considerably small (less than 5 ppm). The largest SCSD value (39.0 ppm) is found in position 1. Both β-positions show the expected negative SCSD values. However, the carbon atom at position 14 is more strongly affected by the hydroxy substituent than the one in position 2. Two structure pairs are shown in Figure 3, which illustrates this point more clearly.

Besides the output of statistical information as shown in Tables I and II, the program can also display the distribution diagrams of SCSD values for each position.[6] The complete SCSD data for each structure pair are stored on a file allowing further detailed investigation.

CONCLUSION

The manual handling of SCSD values of all the possible structure pairs for given substituents from a very large collection of reference data currently available is almost impossible. Based on our efficient MCSS algorithm and some tricky strategies, the SCSD algorithm described here has already proved to be a useful method for automatic extraction of the complete SCSD information from a large database according to one or two input substituents and optionally an additional fragment depending on the user's choice. The program can be used to aid the studies on substituent effects on chemical shifts. Some potential applications of this method have been described in ref 7. The detailed studies of the effects

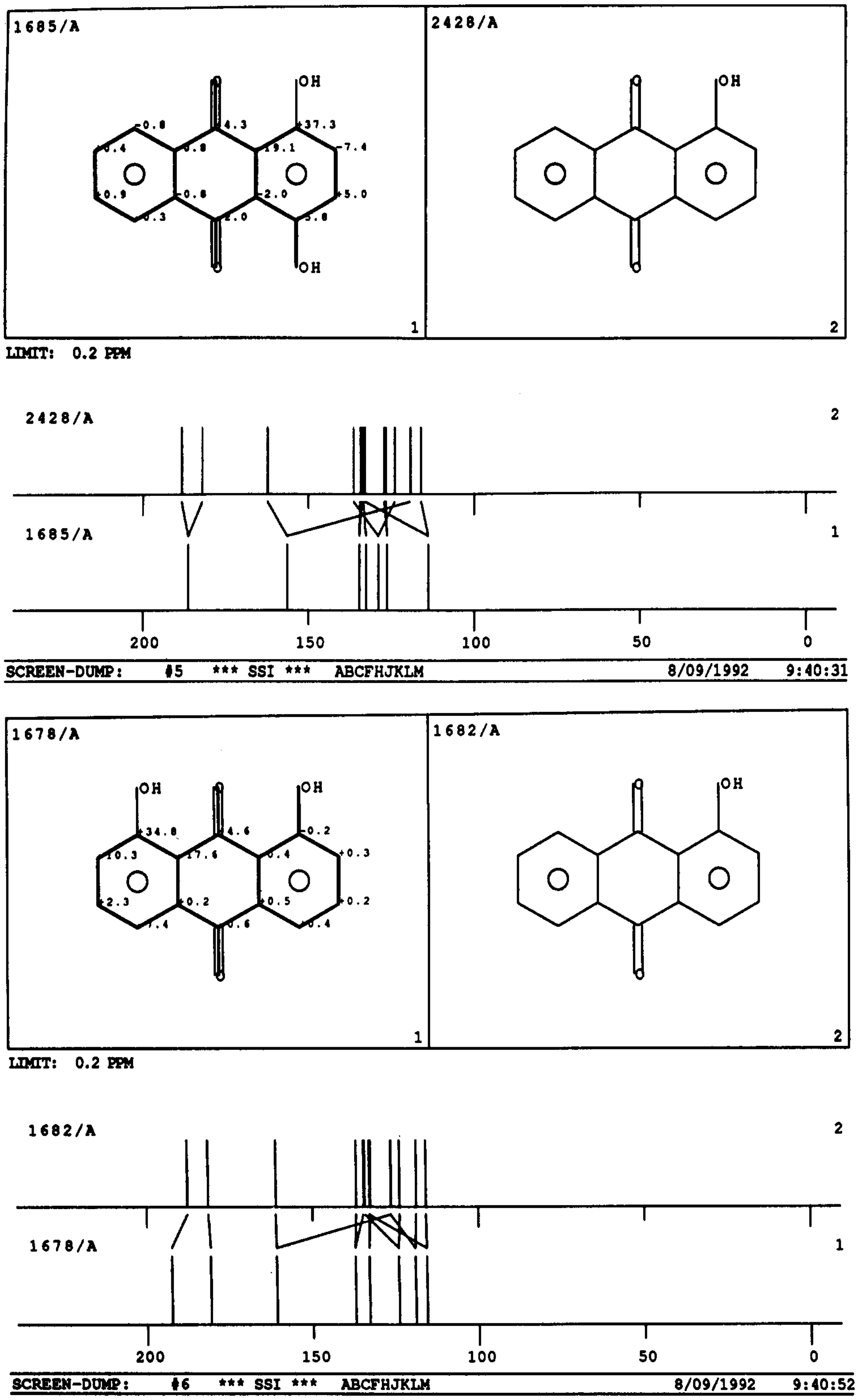

Figure 3. Two structure pairs from Table II. The maximal common substructural fragment (MCSS) is marked by bold lines.

of various substituents on chemical shifts of benzene derivatives by means of the SCSD method are beyond the scope of this paper and will be presented in the following one.[6]

EXPERIMENTAL SECTION

The SCSD algorithm was implemented in FORTRAN 77 under the UNIX operating system on a Silicon Graphics workstation and an IBM-R6000 workstation. The program consists of about 5000 lines of source code. The method has been implemented into the CSEARCH-NMR database system.[8] The databases accessed are the spectral data collections from the University of Vienna, SADTLER Research Laboratories, and the German Cancer Research Center at Heidelberg.

ACKNOWLEDGMENT

L.C. thanks the Austrian Academic Exchange Service for financial support. The authors are grateful to the staff of the University Computing Center for helpful discussion duringthe program development. This project was supported by the European Academic Supercomputing Initiative (EASI).

REFERENCES AND NOTES

(1) Chen, L.; Robien, W. A Novel Approach for Optimized Prediction of ^{13}C-NMR Spectra Using Increments. *Anal. Chim. Acta* **1993**, *272*, 301–308.

(2) Pretsch, E.; Fürst, A.; Robien, W. Parameter Set for the Prediction of the ^{13}C-NMR Chemical Shifts of sp^2- and sp-Hybridized Carbon Atoms in Organic Compounds. *Anal. Chim. Acta* **1991**, *248*, 415–428.

(3) Bremser, W. HOSE-A Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355–365.

(4) Chen, L.; Robien, W. MCSS: A New Algorithm for Perception of Maximal Common Substructures and Its Application to NMR Spectral Studies. 2. Applications. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 507–510.

(5) Chen, L.; Robien, W. MCSS: A New Algorithm for Perception of Maximal Common Substructures and Its Application to NMR Spectral Studies. 1. The Algorithm. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 501–506.

(6) Chen, L.; Robien, W. *J. Chem. Inf. Comput. Sci.*, the following paper in this issue.

(7) Chen, L.; Robien, W. The CSEARCH-NMR Database Approach To Solve Frequent Questions Concerning Substituent Effects on ^{13}C-NMR Chemical Shifts. *Chemom. Intell. Lab. Syst.*, in press.

(8) Kalchhauser, H.; Robien, W. CSEARCH: A Computer Program for Identification of Organic Compounds and Fully Automated Assignment of Carbon-13 Nuclear Magnetic Resonance Spectra. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 103–108.