Paper No. 11

SPECTROSCOPIC DATABASES- A TOOL FOR STRUCTURE ELUCIDATION

P.LUKSCH

Fachinformationszentrum Karlsruhe
D-7514 Eggenstein-Leopoldshafen 2
Federal Republic of Germany

Paper No. 11

SPECTROSCOPIC ONLINE DATABASES- A TOOL FOR STRUCTURE ELUCIDATION

P. LUKSCH
Fachinformationszentrum Karlsruhe, Fed. Rep. of Germany

Spectroscopic databases have developed to useful tools in the process of structure elucidation. Besides the conventional library searches, new intelligent programs have been added, that are able to predict structural features from measured spectra or to simulate spectra for a given structure. The example of the C13NMR/IR database developed at BASF and available on STN is used to illustrate the present capabilities of online databes. New developments in the field of spectrum simulation and methods for the prediction of complete structures from spectroscopic information are reviewed.

Introduction

Spectroscopic data collection in printed form have a long tradition. The first generation of spectroscopic databases offered online [1,2] had obvious advantages compared to the printed collections, but as long as they did not contain fully coded structural information there was a serious limit for the amount of spectroscopic intelligence that can be included in the search or interpretation capabilities.

Within the last decade increased emphasis has been placed on the automatic evaluation of spectroscopic data. New programs, which make use of the correlations between structural and spectral features have been developed to powerful tools in the process of structure elucidation.

Only a part of the features that have been developed in universities or chemical companies are presently available in the online databases. The task of producers of spectroscopic databases should not be restricted to the collection of spectra, but should also focuss on a constant improvement of the interpretation algorithms.

The first part of this paper will describe the present capabilities of online databases, mainly by considering the C13NMR/IR database [3] available on STN. The second part will give an outlook on features

that are in the process of development or which should be considered as potential improvements.

1. Present search capabilities

At present the STN database contains about 100,000 C13NMR spectra and 16,000 IR spectra. The structures are coded as connection table and they can be displayed in text format or in graphical form as shown in Fig.1. The NMR spectrum has been reconstructed from the peak list. An intensity of one unit corresponds to the signal of one carbon.

Only the aspects that are typical for spectroscopic files will be discussed here. Search capabilities for substance information by chemical names or molecular formula are also available in other kind of substance-oriented files and can be considered as standard search techniques today.

1.1 Search for spectral information

A typical query is to find out wether the reference database contains spectra which are similar (or identical, which is the easiest case) to the measured spectrum of the unknown compound. For this purpose various search techniques are available:

In the single peak search the user enters a sequence of peaks from his query spectrum with associated information like intensities or multiplicities for NMR spectra. The search result will consist of spectra containing the query peaks within the user-specified tolerance. The disadvantage of this method is that a database spectrum will be rejected if only one peak is missed, although it could well be considered to be "similar" to the query spectrum.

More sophisticated algorithms have been developed for similarity searches. In this type of searches, the query is the complete spectrum and it is run against every (except if a presearch has been executed) spectrum of the database in a sequential way.

Spectroscopic Knowhow has been included in these programs and therefore each spectroscopic method needs its own algorithm. Deviations between query spectrum and reference spectra will be accepted as long as the overall match or similarity surpasses a certain threshhold value.

Fig.2 shows the first four hits of a similarity search with infrared spectra.

If the user is interested in seeing the similarity of the structures found in the search, he can also display the structure images of the best matching structures as shown in Fig.3.

In some cases, a large amount of structures can be found. Therefore, good ranking algorithms are required to display the results. The SISCOM search for mass spectra [4] for example contains a useful feature. In order to facilitate the overview of the relevant structures, very similar structures - for example several derivatives of the same compound - are shown only once in the hit list.

However, even with the best search algorithm, valid results can only be obtained when similar spectra or structures are present in the database.

1.2 Spectrum Interpretation

For C13NMR, interpretation rules have been derived from the database. A set of 1500 partial structures or transcripts with assigned CNMR and HNMR shifts have been coded, mainly manually. Whenever the spectral pattern of such a transcript occurs in a spectrum, there is a high probability that this partial structure is present in the unknown struture. In order to improve the selection of correct transcripts, restrictions on the elemental composition of the structure can be specified.

As an example, the C13NMR spectrum of n,n'-dimethyl-4-nitro-benzalhydrazone chloride (shown in Fig.1) has been entered with the additional restriction that no sulfur should be present in the structure. Nine possible transcripts have been proposed, and as shown in Fig.4, two partial structures, the nitro-phenyl (Trc.1244) and the dimethylamino group (Trc.1344) have been correctly recognized by the interpretation algorithm.

1.3 Spectrum estimation

The C13NMR chemical shift is correlated with the chemical environment of the carbon atom. With the HOSE codes developed by Bremser [5], a method for the automatic generation of the mean resonance frequencies of all chemical environments from a large database has been introduced.

This method works with a very good precision for C13NMR. Attempts to simulate other kind of spectra will be discussed later.

In many cases, the chemist has already a proposal for his structure. The spectrum estimation is an easy way, either to confirm the postulated structure or to get an indication for which parts of the structures the estimated and measured chemical shifts are in disagreement.

Fig.5 shows how a query structure is built up with simple key-boarding commands. The bottom window of the screen contains the list of the available fragments that can be used to model a structure. Any structure of the database can also be loaded and modified.

For each carbon atom the multiplicity (determined from the structure), the average shift with standard deviation and the number of measured shifts involved in the calculation are displayed. If no carbon atom with identical HOSE code (i.e. identical chemical environment) is found, the shift value is obtained by interpolating between the values from the two closest HOSE codes. For carbon C6, two estimated values of 133.7 and 127.4 are given. Due to the limitation of four spheres for the environmental code of C6, the influence of the Fluorine atoms cannot be taken into account in the estimation. However, the program has recognized that the measured values for the required code are grouped around two maxima. In order to decide which value is the correct one, the user has to retrieve the original structures and the measured shifts which have been used for the estimation. This will show him that the higher value has to be assigned to C6.

2. Future Developments

The features that have been discussed so far, are available online on STN. Additional developments have been performed or will be done in the next years in the field of the database production and of new software tools.

2.1. Database production

The Government in the Federal Republic of Germany supports the development of an integrated spectroscopic information system which includes spectra of various spectroscopic techniques. Besides the relatively large existing collections of C13NMR, IR and Mass spectra, it is planned to buid up smaller pools of NMR spectra of other nuclei (H, N, P, O) and of UV and Raman spectra. The whole system will consist of a central structure file with pointers to the different spectral files.

2.2. Software Developments

The mentioned spectral pools build the knowledge base for the Specinfo expert system developed in cooperation with BASF. The aim is to extend the interpretation rules to the other spectroscopic methods. For Infrared spectra, statistical methods have been used to derive correlation rules between subspectra and substructures.

Other groups cooperating in the field of mass spectra are using different approaches:

- At the University of München, a program for the prediction of mass spectra is under development [6,7]. In this model chemical and physical properties such as dissociation energy, charge distribution and polarizability are calculated and used for prediction of the fragmentation process. In addition another approach, where the relation between spectrum and structure is described in an associative memory system is also followed.

- The EDAS system [8] is an interactive program which combines the intellectual knowhow of the spectroscopists and the ability of the program to recognize substance classes from the measured spectrum.

Each spectroscopic method is sensitive to different structural features and the aim of the Specinfo system is to combine all interpretation rules for the generation of complete structures [9,10]. Spectrum prediction algorithms can then be used either to confirm the candidate structures or to reject substructures.

Specinfo is designed as an inhouse system (presently for VAX-VMS systems). It allows the user to improve the knowledge base by adding his own spectra and structures to the database.

2.3. STN Implementation

FIZ Karlsruhe is working on a new implementation of spectroscopic files in the STN Messenger retrieval software. It is planned to add as many features of Specinfo as possible to the online system. The disadvantage of missing the newest developments of Specinfo will probably be compensated by the large amount of spectral files available in the online version. It is also planned to have an STN-wide consistent substance identification, which will permit an easy crossover to other files like Registry or Beilstein.

References

[1] HELLER, S.R., EPA/NIH Mass Spectral Database, US Gov. Printing Office, Washington (1978).
[2] NIH/EPA Chemical Information System, User support, Falls Church, VA 22046.
[3] Carbon 13 Nuclear Magnetic Resonance and Infrared Database, STN International, Federal Republic of Germany.
[4] MS Online, Mass Spectra Database, Fachinforma tionszentrum Chemie, Postfach 126050, Berlin
[5] BREMSER, W., Anal. Chim. Acta 103, (1978) 355
[6] HANEBECK, W., SALLER, H., GASTEIGER, J., in J.Gasteiger (Hrsg.): Software-Entwicklung in der Chemie 2. Springer-Verlag, Berlin (1988) 197
[7] GASTEIGER, J., Nachr. Chem. Tech. Lab. 34 (1986) 226
[8] VARMUZA, K., WERTHER, W., LOHRINGER, H. in G. Gauglitz (Hrsg.): Software-Entwicklung in der Chemie 3. Springer-Verlag, Berlin (1989) 267
[9] BREMSER, W., FACHINGER, W., Multidimensional Spectroscopy, Magn. Reson. Chem. 23 (1985) 1056
[10] SCHUBERT, V., BREMSER, W., NEUDERT, R., KUBINYI,H. Der Computer in der Massenspektrometrie, Nachr. Chem. Tech. Lab. 37 (1989) 720

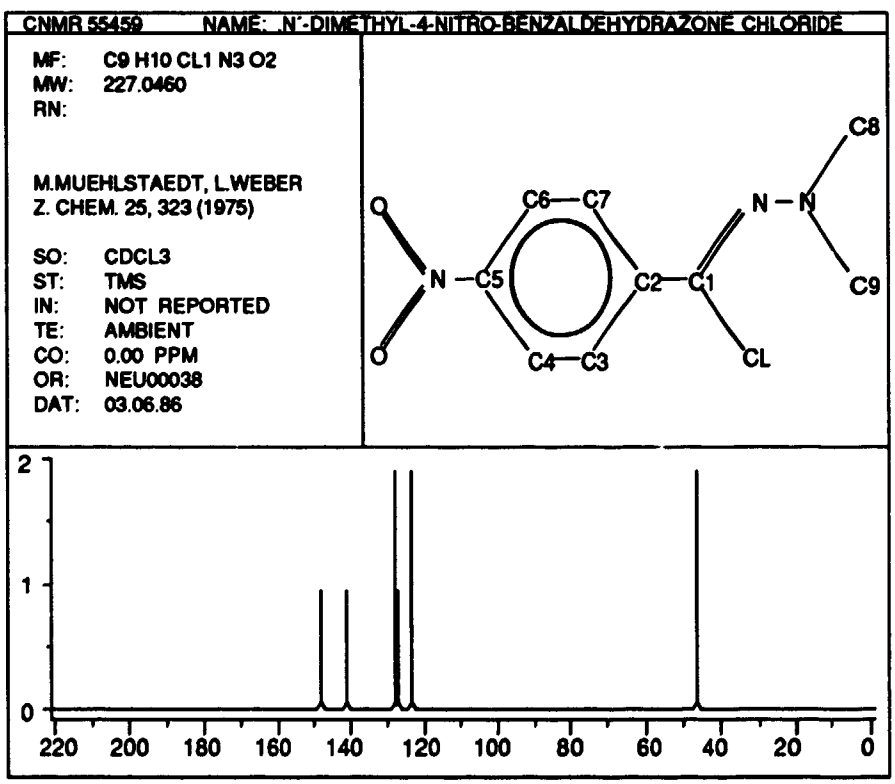


Fig.1 C13NMR sample record in graphic mode

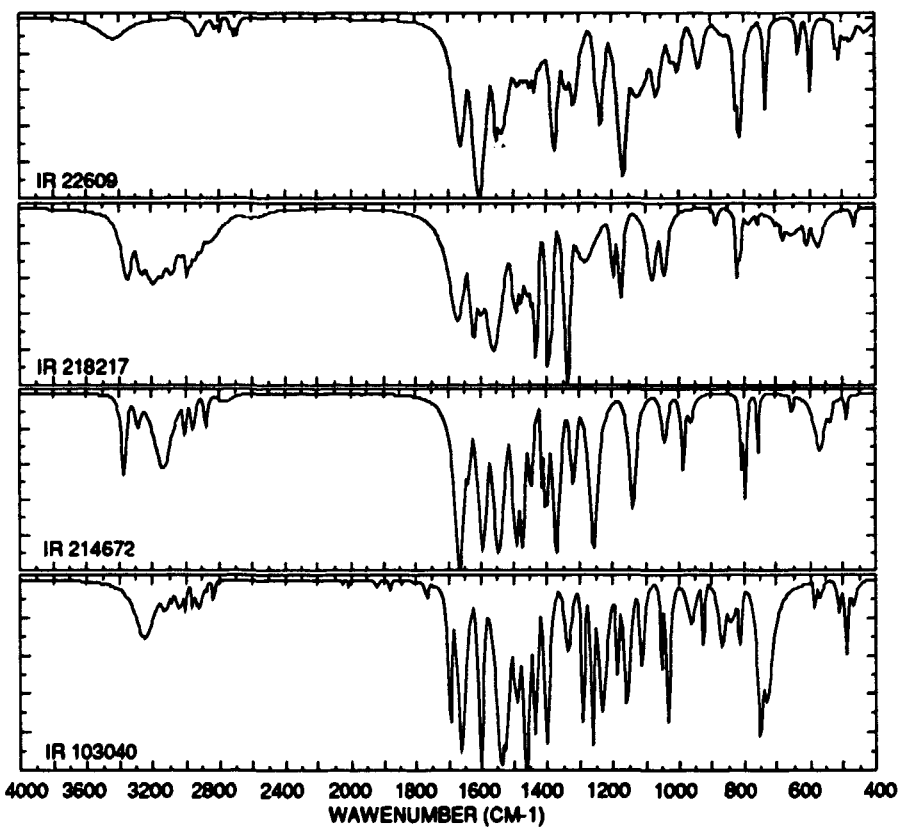
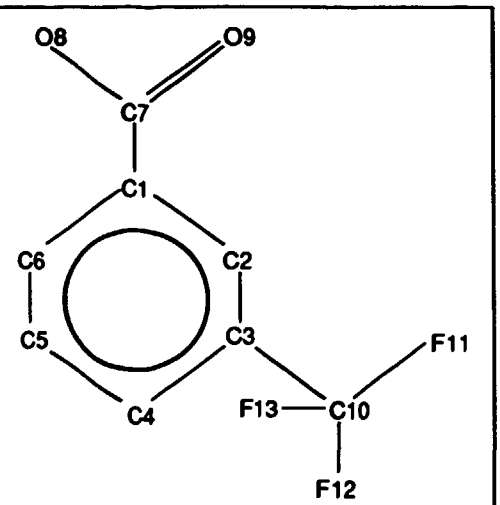


Fig. 2 Four best matching spectra from similarity search

OPTIONS: > BENZOH,1.COOH,3.CF3,P,S

ESTIMATED SHIFTS (PPM):

C1	S	132.7	+-0.0	(1	LI)
C2	D	126.7	+-0.1	(4	LI)
C3	S	130.7	+-1.1	(17	LI)
C4	D	130.3	+-2.0	(12	LI)
C5	D	129.5	+-0.2	(2	LI)
C6	D	133.7	+-0.2	(3	LI)
		127.4	+-0.6	(12	LI)
C7	S	168.4	+-1.9	(30	LI)
C10	S	123.9	+-0.3	(15	LI)



K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
R3	R4	R5	R6	R6L	R7	R8	R9	R10	R11
R12	DECALI	BENZOL	BENZOH	BENZOS	PYRIDI	PIPERI	PIPERA	PIRIMI	MORPHO
ANDROS	NORBOR	ADAMAN	CHOLES	NAPHTH	ANTHRA	RETINA	INDANE	TETRAL	PYRROL
PYRAZO	IMIDAZ	TRIA-A	TRIA-S	FURAN	THIOPH	INDOL	INDA-H	TROPYL	I3
S4	I4	T4	NO2	CF3	CCL3	COOH	COOCH3	COCH3	CHO
ACETOX	OCH3	OC2H5	CH2OH	C2H4OH	SO3H	VINYL	ALLYL	CN	ISO-CN
NCO	COOC2H	CYPED-	BIPHEN	BENZOA	TETRA				

Fig. 2 C13NMR spectrum estimation

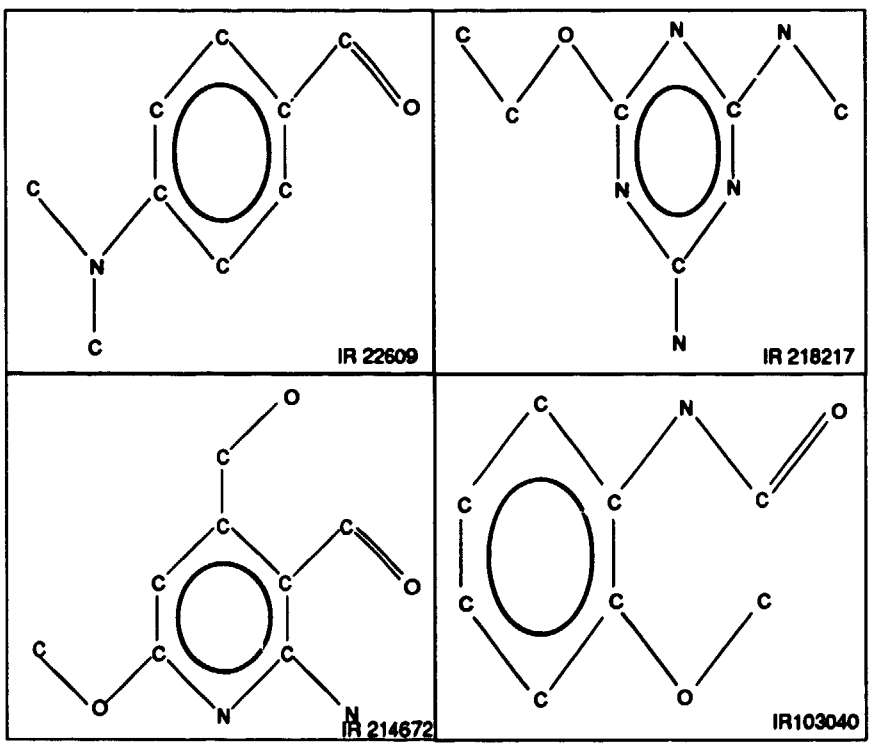


Fig. 3 Four best matching structures from similarity search

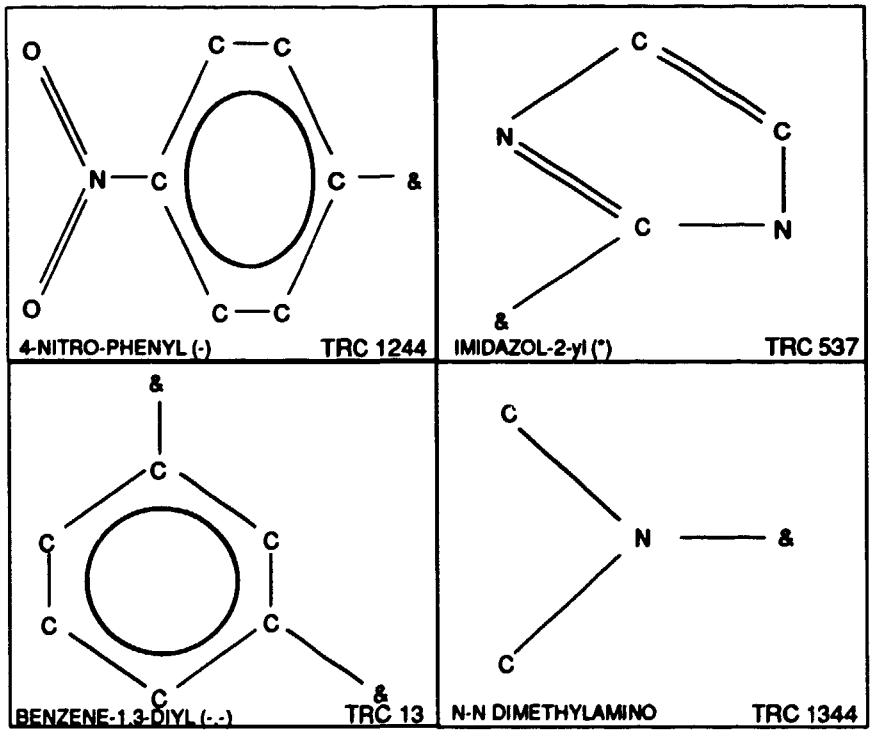
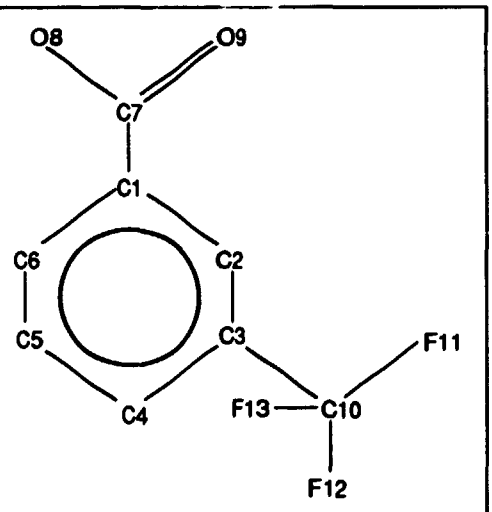


Fig. 4 Best four substructures proposed in spectrum interpretation

OPTIONS: > BENZOH,1.COOH,3.CF3,P,S

ESTIMATED SHIFTS (PPM):

C1	S	132.7	+-0.0	(1	LI)
C2	D	126.7	+-0.1	(4	LI)
C3	S	130.7	+-1.1	(17	LI)
C4	D	130.3	+-2.0	(12	LI)
C5	D	129.5	+-0.2	(2	LI)
C6	D	133.7	+-0.2	(3	LI)
		127.4	+-0.6	(12	LI)
C7	S	168.4	+-1.9	(30	LI)
C10	S	123.9	+-0.3	(15	LI)



K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
R3	R4	R5	R6	R6L	R7	R8	R9	R10	R11
R12	DECALI	BENZOL	BENZOH	BENZOS	PYRIDI	PIPERI	PIPERA	PIRIMI	MORPHO
ANDROS	NORBOR	ADAMAN	CHOLES	NAPHTH	ANTHRA	RETINA	INDANE	TETRAL	PYRROL
PYRAZO	IMIDAZ	TRIA-A	TRIA-S	FURAN	THIOPH	INDOL	INDA-H	TROPYL	I3
S4	I4	T4	NO2	CF3	CCL3	COOH	COOCH3	COCH3	CHO
ACETOX	OCH3	OC2H5	CH2OH	C2H4OH	SO3H	VINYL	ALLYL	CN	ISO-CN
NCO	COOC2H	CYPED-	BIPHEN	BENZOA	TETRA				

Fig. 5 C13NMR spectrum estimation