WinDat: An NMR Database Compilation Tool, User Interface, and Spectrum Libraries for Personal Computers

Sergei V. Trepalin,* Alexander V. Yarkov, Ludmila M. Dolmatova, and Nikolai S. Zefirov

Institute of Physiologically Active Compounds, 142432 Chernogolovka, Moscow Region, Russia

Simon A. E. Finch

Chemical Concepts GmbH, D-69469 Weinheim, Germany

Received September 14, 1994[⊗]

A fast, efficient program which runs on the ubiquitous personal computer for the analysis, storage, and retrieval of NMR information is presented. It is suitable for abstracting large quantities of data from published literature and includes many plausibility tests which are executed simultaneously with the input. Automatic determination of stereoconfigurations in 2D structures makes the program of great value for natural compounds. A structure elucidation system allows prediction chemical shifts and *JJ* coupling constants and automatic peak assignation. A database of 101 205 ^{13}C, ^{31}P, ^{19}F, ^{29}Si, ^{15}N, ^{11}B, ^{17}O, and ^{33}S spectra, taken primarily from Russian-language original sources, has been created.

INTRODUCTION

The software currently available to support identification of chemical substances from the data of nuclear magnetic resonance (NMR) measurements comprises database systems such as SpecInfo[1,2] and CSEARCH[3] which use algorithmically-encoded empirical knowledge to abstract information from a database of atom-centered structure-spectrum relationships and structure elucidation programs such as CHEMICS[4] and X-PERT (formerly RASTR),[5,6] though these have the disadvantage of requiring too much information about the structure sought (i.e., its brutto formula or structure, respectively) to have wide application. Developments now in hand include attempts to couple structure elucidation with database software (SpecInfo/CHEMICS,[6] SpecInfo/MOLGEN,[6] WIN-SpecEdit,[6] and SpecTool[6]) to combine their strengths and work on neural networks with the objective of obviating the need for the database.

The database nevertheless remains a central requirement either directly, as in the case of SpecInfo and CSEARCH where it constitutes part of the commercial product, or indirectly, to constrain the structure generation modules or as a source of training sets for the neural networks. Consequently there is a requirement for software which can be run on machines as small as a notebook PC for use in abstraction of literature data and/or documentation of laboratory results with the capability of searching for and predicting spectra. Here we describe WinDat,[6] the outstanding features of which are extensive automated checking of input data, particularly the correct assignments of chemical shifts and *JJ* coupling constants, to support abstraction/documentation (since experience shows this to be a critical step as errors made at this time are very difficult to correct later) and a compilation of over 100 000 spectra of different nuclei.

PROGRAM DESCRIPTION

1. Program Architecture. WinDat is an interactive, menu-driven program, which runs under the Microsoft Windows 3.1 operating system. Its architecture is shown schematically in Figure 1. The database consists of a number of data structures, which we shall call *files*. A user can create a new file or delete an existing one by selection of the appropriate commands from a menu. Each file is used to store NMR data for a particular nucleus and currently ^{13}C, ^{31}P, ^{19}F, ^{29}Si, ^{15}N, ^{11}B, ^{17}O, and ^{33}S are supported though there is no principle problem to expand the list. A special remark about ^{1}H spectra is appropriate. It is possible to represent these spectra as peak tables containing chemical shifts and *JJ* coupling values, but in the published literature data on ^{1}H spectra are usually incomplete and cannot be used for spectrum simulation. The usual way to store ^{1}H NMR data is to use the full digitized spectrum.

Each file consists of a number of records comprising the NMR spectrum peak table, its assignment, chemical structures, and auxiliary information. Records may be copied from one file to another or duplicated within a file *via* the clipboard. The user can create, delete, and modify records.

Input of numeric and text data for a record is performed in the usual way for Windows applications and will be not discussed here.

2. Chemical Structure Input. A typical feature of a chemical database is the presence of chemical structures stored as connection tables. WinDat can handle structures of 255 atoms and bonds and will normally calculate attached hydrogen atoms "on-the-fly" except where they are explicitly defined to code stereo information. Aromatic structures are recognized automatically by application of the Hückel rules. In common with many other computer graphics programs for chemical applications, WinDat contains a structure drawing editor which makes extensive use of mouse buttons and keyboard duplication of commands which are summarized in Table 1. Some of them perform actions which will be familar from other PC programs, so we shall discuss here only those aspects which are new or are significant improvements on existing structure-editing software.

Glossary. Reference 7 contains the first published and fully assigned ^{13}C spectrum of 26-chloro-26-deoxycryptogenin, but an image of the structure is omitted. This is often

[⊗] Abstract published in *Advance ACS Abstracts,* April 15, 1995.

0095-2338/95/1635-0405$09.00/0

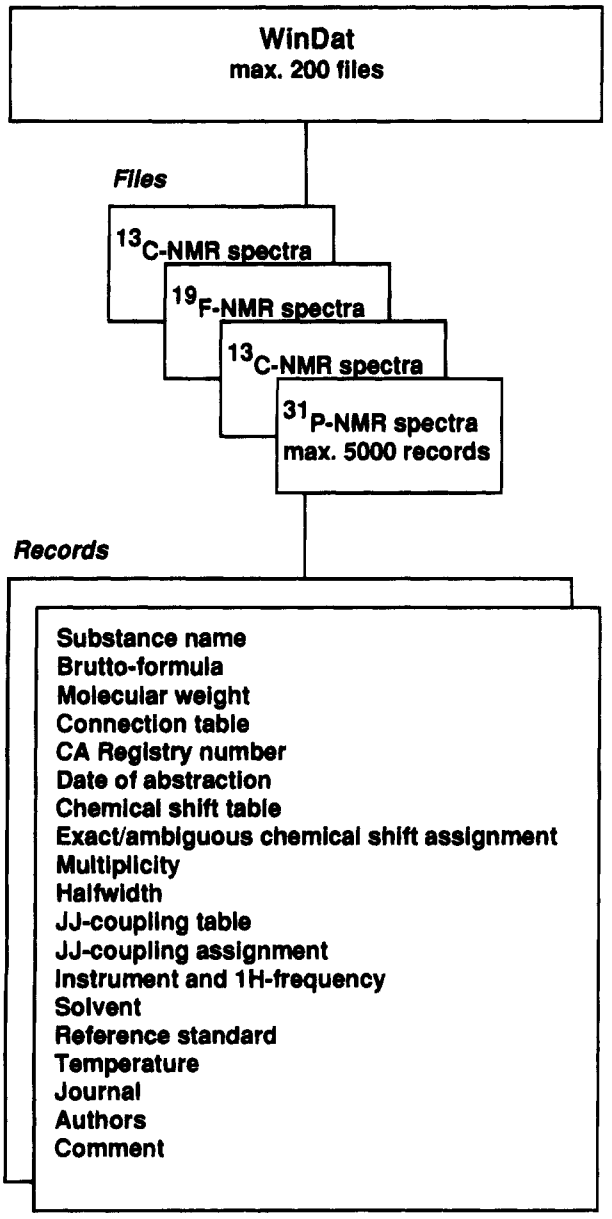

Figure 1. Architecture of WinDat.

the case with previously documented natural compounds. Using the glossary one can frequently avoid the labor of referring to other literature. The glossary is a predefined database which contains trivial names and associated structures with commonly used atom numbering. Stereo information is inserted where possible. Its current content is 17 000 records, and it is continuously expanded. It covers almost all of the trivial names used in general organic chemistry and many of those used in the chemistry of natural compounds, biochemistry, and pharmacology. To extract a structure from the glossary it is necessary to define a substring of it, after which a search is made and the program proposes a selection of the names found (Figure 2). Selection of a name from the list causes the associated structure to be loaded.

Stereonotation. A remarkable feature of natural compounds is the existence of a large number of isomers. For example, hexopyranose has five asymmetric centers and hence 32 stereoisomers. There are some commonly used ways to represent a three-dimensional structure in two dimensions of which some examples are shown in Figure 3. To unify all three structures stereo labels (*R/S/Z/E*) should be assigned to appropriate atoms and bonds. In most cases the stereo labels are not defined explicitly in the original publication, so the appropriate calculations applying the Cahn–Ingold–Prelog (CIP) rules must be done. Like the well-known CHIRON[8] program, WinDat recognizes stereocenters drawn with "up" or "down" stereochemical bonds in 2D structure images and Fischer projections in accordance with the CIP rules and will also generate and display such bonds on the molecular graph if an *R* or *S* label is put on an atom. In addition (a feature lacking in CHIRON), a new type of stereo structure—cyclic sugars—is also recognized (Figure 4). This last type of stereo structure is the most complicated for analysis and in the current version the program makes erroneous assignments in some cyclic structures which lack stereo information. Manual correction of stereolabels is possible in such cases. The program has predefined list of stereo labels—*R*, *S*, *R or S*, *Z*, *E*, *Z or E* (Table 1). Unlike CHIRON, 3D structures are not accepted by WinDat.

Template Windows. Up to four additional windows with user-defined templates may be opened simultaneously with the main window and activated by mouse movement inside the visible area at any step of structure creation.

3. Spectroscopic Data Input. The values and assignments of chemical shifts and *JJ* coupling constants are the major content of an NMR database. Some special routines are required to assign chemical shifts or *JJ* coupling constants to atoms or pairs of atoms, respectively. Input of data which is specific for NMR is discussed here; input of all other data is performed in the usual way for Windows applications.

Assignment of Chemical Shifts and *JJ* Coupling Constants. Published data frequently contain only partial assignment of chemical shifts. There are two reasons for this: some problems solved by NMR spectroscopy do not require complete assignment and in these cases a human expert can assign the remaining chemical shifts. The second (and most frequent) reason is the impossibility of precise assignment of all chemical shifts on the basis of an NMR experiment where a number of magnetically nonequivalent atoms may be assigned to a given chemical shift. For these reasons two types of assignment—*exact* and *ambiguous*—have been introduced.

JJ coupling constants must be assigned to pairs of atoms. This is impossible if the corresponding chemical shifts have been assigned ambiguously, and we have not yet satisfactorily resolved this technical dilemma. One work-around is to link the *JJ* coupling to the chemical shift at which it is measured, but this is of little use in structure elucidation and makes no use of the fact that the type of atoms interacting and the order of the ambiguous *JJ* coupling constant are frequently known.

Multiple Reference Standards. There is no difficulty about the reference standards used to define chemical shifts in ^{13}C and ^{31}P spectra because TMS and H_3PO_4 are used for ^{13}C and ^{31}P spectra, respectively, in more than 99% of the published data (P_4O_6 is recommended by the ASTM committee for ^{31}P chemical shifts measurements,[9] but this standard has effectively been ignored). On the other hand, a number of reference compounds are used for ^{19}F, ^{11}B, ^{15}N, ^{29}Si, and ^{33}S spectra so it is necessary to recalculate all chemical shifts to a selected reference compound for solving

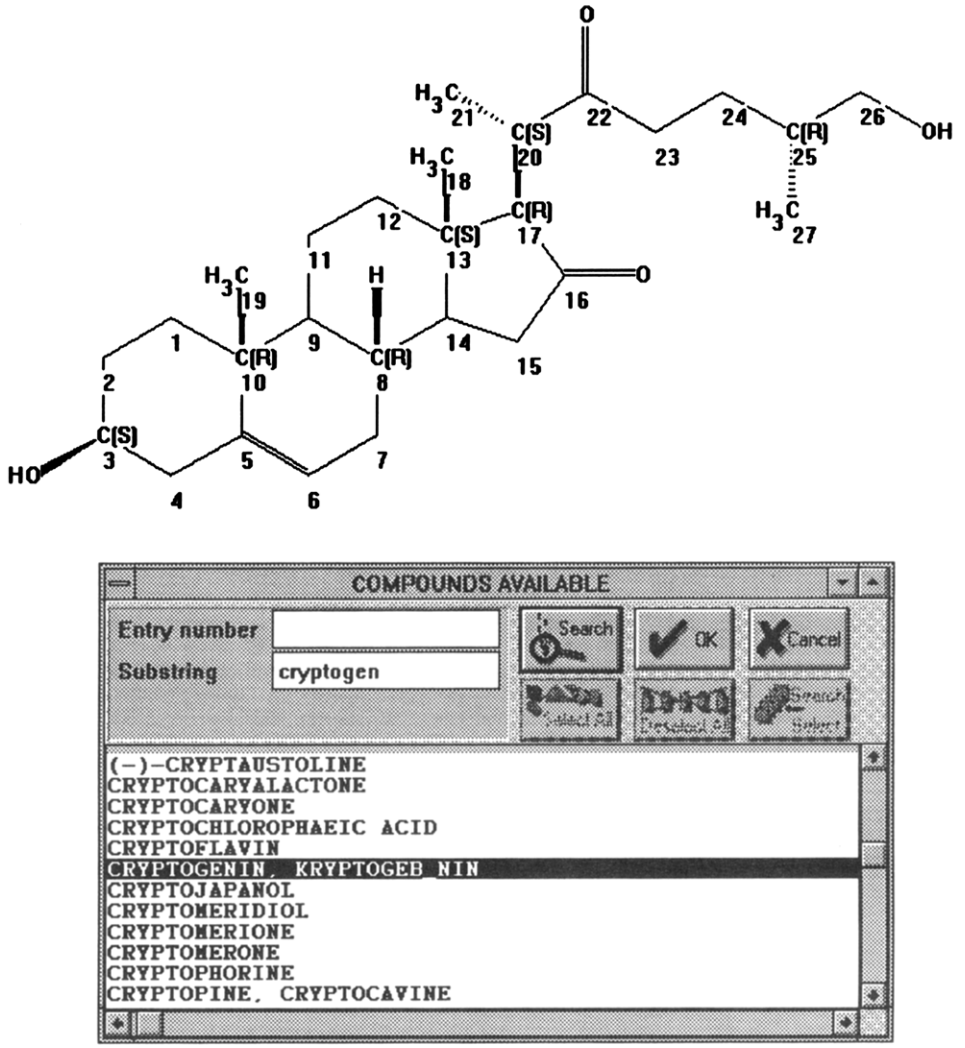

Figure 2. Extraction of a record from the glossary.

Table 1. Summary of Commands for the Structure Editor

command	action/description
atoms	H, C, N, O, P, S, F, Cl, Br, Si, D, others from the periodic table and amino acids
atomic attributes	charge, isotope, valency, radical, biradical
bonds	single, double, triple, up, down, either, coordinate
copy/paste	clipboard operations
delete	deletes atoms, bonds, groups, fragments, unconnected atoms
draw	normal, rubberband, grid, cycle and chain
glossary	displays structure with atomic numbering for a defined trivial name
invert	inverts *x*-coordinate, *y*-coordinate, exchanges groups connected to common atom
load	internal format, MOLFILE, JCAMP, SMD
move	moves fragments, groups, atoms, and bonds
options	colors, fonts, display implicit H, auto-fit on window-edge overlap, flip elsewhere, redraw bonds ratio
query atoms	hetero, halogen, metal, any, exactly-coordinated
query bonds	any, aromatic, single or double, chain, ring, cis/trans
redraw	redraws structure, fragment group or bonds
resize	draws structures and bonds larger or smaller, scales to full screen or "like bond"
rotate	rotates fragment or group, flip, bond horizontal or vertical, about *Z*-axis
save	internal format, MOLFILE, JCAMP, SMD, WMF
stereoconfiguration	determines *R/S*, *Z/E*, all stereocenters in a structure, applies labels *R*, *S*, *R or S*, *Z*, *E*, *Z or E*
templates	rings, stereorings, condensed rings, chain/ions, chemical groups, user-defined, use screen contents
undo	reverses up to three most recent commands

structure elucidation problems and predicting chemical shifts. At the same time, it is desirable to have chemical shifts as they were presented in the original for verification of data. In WinDat the problem has been solved by creating lists of reference compounds for each nucleus of interest in which each reference compound has its chemical shift value defined relative to the major standard: $CFCl_3$ for ^{19}F spectra, $BF_3(OEt)_2$ for ^{11}B, CH_3NO_2 for ^{15}N, TMS for ^{29}Si, H_2O for ^{17}O, and CS_2 for ^{33}S. All chemical shifts are entered and stored as they are defined in the original reference. During calculation of the database used for chemical shift prediction they are all referred to the major standard, but the user may select any desirable reference compound when looking at chemical shifts and their assignments.

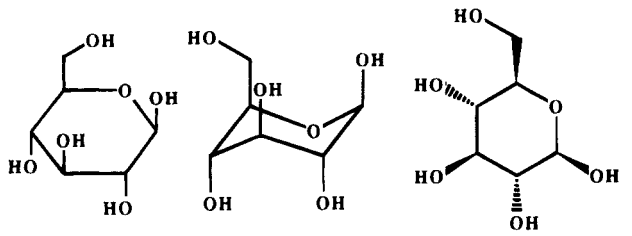

Figure 3. Three representations of β-D-glucopyranose.

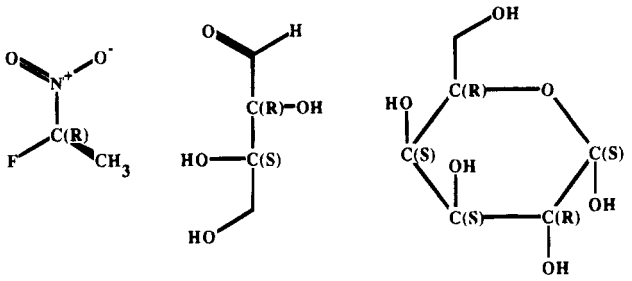

Figure 4. Type of structures for which automatic calculation of stereoconfigurations is possible.

Bibliographic Information. WinDat employs separate fields for authors, journal title, volume number, first and last pages, and year of publication. Such division enables a number of actions which are not readily possible if the reference is stored as a string, in particular (i) locating publications to verify data, search for data from a given article, and check chemical shift direction automatically (see below) and (ii) determining the number of publications used as a measure of the confidence level when predicting the chemical shifts of fragments—many references to a single publication does not have the same confidence level as the same number to many publications.

Use of Data Stored as Lists. Some supplementary information in the NMR database—solvents, reference compounds, journal titles, names of instruments—are represented as string data in fairly short indexed lists with only the index key associated with each spectrum record. This has the twin advantages of conserving disk storage and ensuring standardization of spelling and thus completeness in searches. The user is relieved of the burden of having to remember the keys by being able to select appropriate items from a list which he can expand at will. Standardization has significant value for particular purposes such as investigation of the influence of solvents on chemical shifts, precise spectrum simulation (when the ^1H frequency of instrument must be known), etc. The usual drawback of this method is the incompatibility of data entered at different work places, and to overcome this WinDat has a routine which compares data from two sources and creates a common list and database.

Interfaces to Other Programs. The JCAMP format[10,11] is used to exchange data with other programs. A high-quality structure image may be saved in a placeable Windows Metafile which can be read by some Windows applications (e.g., Microsoft Word). Different formats are available for import and export of bond-connection matrixes (Table 1).

4. Verification of NMR Data. Chemical shifts and *JJ* coupling constants have the property of transferability, that is, they are similar for identical fragments in different compounds. This property is the basis of NMR data verification. All the tests described below (excluding chemical shift axis direction) are performed automatically as a new record is added to the database. All errors and warnings are brought to the user's attention immediately so they may be dealt with while the source literature is at hand. The time required for all tests does not exceed 6 s in the worst case (^{13}C spectra, where nearly 5 Mbyte are read from hard disk). The tests may also be performed on the whole content of the database.

Database for Chemical Shift and *JJ* Coupling Constant Prediction. The first step in calculation of the database for chemical shift and *JJ* coupling constant prediction is the collation of statistical information (average, standard deviation, minimum, maximum, and number of observations) for each structural fragment. WinDat uses an atom-centered substructure code like HOSE[12] to describe structure fragments in terms of concentric "spheres" of neighboring bonds and atoms, but unlike HOSE each descriptor is a 4-byte hash value. For the first sphere this is generated by coding information on the atom type (5 bits), charge (2 bits), valency (2 bits), cycle size and type (3 bits), and bond type (2 bits) into a 14 bit variable for each neighboring atom and then bit-shifting and XORing these values to generate a descriptor in the form of a pseudo-random number derived from the physical environment. Descriptors for subsequent spheres are generated by bit-shifting and XORing first sphere descriptors centered on the atoms of the outer spheres. Compared to HOSE this scheme has the advantages of constant descriptor length and of including ring information for cyclic fragments in the first-sphere descriptor. Its disadvantages are that the function cannot be inverted to obtain information on the chemical structure from the descriptor, and because *n*-sphere codes are not subsets of *n* + 1 sphere ones it cannot be used to interpolate chemical shift values for unknown descriptors. (HORD,[12] a modified HOSE code which is used for coding cycles, is not centered on an atom and hence cannot be used to predict chemical shifts and *JJ* coupling constants.) Stereo information is not included in the hash function and consequently plays no role in assignment or prediction of chemical shifts. It is possible for the same descriptor to be generated for different environments. However, we have not detected such events for first sphere environments, and it has occurred in only 2 out of 50 000 second sphere environments with the standard deviation of the predicted chemical shift being raised significantly in both cases.

To add new information about chemical shifts (bulk loading of the database), WinDat reads files of coded fragments and their statistical data and then reads the new structures. Prior to descriptor generation a number of modifications are applied to input structures: (i) all explicitly-defined hydrogens are eliminated, (ii) all semipolar bonds are converted to double bonds to avoid generation of different descriptors for the same fragment written in different ways (as, for example, O=N=O and $O=N^+ -O^-$), and (iii) all stereo bonds are substituted by single bonds. For each exactly assigned chemical shift descriptors are generated starting from the first sphere and working outwards to a user-controlled maximum of two to five spheres. If a descriptor already exists in the database, its statistical data are changed, otherwise it is inserted.

Fluorine is always univalent, and the sphere for descriptor generation is centered on the neighboring atom to remove redundant first sphere data.

Calculation of new database entries for *JJ* coupling constants differs from that for chemical shifts. First, if a hydrogen atom is involved in the interaction, it is not removed from the structure. Two descriptors centered on the interacting atoms and using the types of atoms under interaction and the order of the *JJ* coupling constant are used as the first sphere are generated and then XORed to construct the final descriptor.

Verification of Chemical Shifts and *JJ* Coupling Constants. The plausibility of chemical shifts and *JJ* coupling constants is checked by comparing the input values with those already in the database, starting with the first-order sphere and working outwards until there are no higher-order spheres or the difference between the input and the database values exceeds three standard deviations. In the latter event a warning message is generated. The program also checks (i) that each atom (of the type recorded in the file) is assigned a chemical shift, (ii) that each chemical shift is assigned to an atom, (iii) that no atom is simultaneously assigned exactly and ambiguously, and (iv) that no atom remains unassigned while another atom is assigned more than once. Detection of any of these suspect conditions causes a warning message to be generated.

Verification of Chemical Shift Axis Direction. The ASTM committee requires that the scale to high frequency (low field) from the reference signal shall be positive.[9] Unfortunately, this convention has not been followed throughout the history of NMR spectroscopy, and furthermore it is not adhered to by some chemists today. Worse still, the sign convention is not explicitly given in more than 30% of original publications. This can confuse even an expert operator when the chemical shifts of the compounds under study are close to zero relative to the chosen reference as happens, for example, when ^{31}P spectra are recorded for some monophosphates relative to H_3PO_4 in the absence of other compounds which would provide orientation.

The procedure for chemical shift axis verification can be executed only for compounds which are already recorded in the database. The square difference of the published to the database values is calculated. The candidate chemical shifts are then sign-inverted, and the comparison is repeated. If the difference of candidates to database is smaller in the second case than in the first, a warning message is generated, and WinDat automatically inverts the candidate chemical shifts if the user confirms the action. (For greater confidence this operation may be performed on a list of the compounds in a given publication rather than on each compound individually.)

Verification of Chemical Structures and Compound Names. Potential errors in chemical structures and compound names are detected by a succession of filters which generate warning messages. This is the case if a match is found in an (editable) ASCII list of common misspellings of substrings. Expected correlations between the brutto formula and the compound name are also checked: for instance, fluorinated compounds have the substring "fluor" in their name and sulfur is frequently associated with the substrings "thi", "sulf", and "mercap". The reverse is also true, so if a name contains the substring "phosph" there is a high probability that the compound contains phosphorus. (However, the user should be aware of the limitations of this technique as exemplified by the nonoccurrence of fluorine in fluorene.) As with the misspellings list, the list correlating atomic symbols and substrings can be edited.

Table 2. Summary of Search Commands

command	description
molecular weight	numeric search
temperature	numeric search
exact structure	search for exact match with query
fragment	substructure search—see text for details
peak histogram	NMR-specific search—see text for details
year of publication	numeric search
volume of source	numeric search
source number	numeric search
compound name	substring search
author's name	substring search
comment	substring search
solvent	search for exact match with query
reference standard	search for exact match with query
instrument	search for exact match with query
source name	search for exact match with query
brutto formula	the lower and upper limits of the formula are defined; a list of elements which must be present may be specified
CA registry no.	search for exact match with query
peak table	inclusion of query in a spectrum
halfwidth of peaks	numeric search
JJ coupling value	numeric search
JJ coupling of labelled atoms	NMR-specific search—see text for details

5. Interpretation of NMR Spectra. Much of the value of the data abstracted and verified by the software described above is its use in structure elucidation: WinDat contains routines for prediction of chemical shifts, *JJ* coupling constants and peak assignment. Search routines are available for all types of records, as summarized in Table 2. Case-sensitive and -insensitive modes may be used in substring searches, equivalent, inside and outside intervals, and greater and less than operators in searches of numerical data. Searches may be continued with AND, OR, and NOT logical operators, stored, recalled, or discarded to restore the previous state.

Prediction of Chemical Shifts and *JJ* Coupling Values. Chemical shifts and *JJ* coupling constants for fragments, calculated as described above (see Database for chemical shift and *JJ* coupling constant prediction), are used to predict the spectra of query structures. The prediction includes assignations and chemical shifts with minimum and maximum values, standard deviation, total number of observations, and size of the spherical environment for the relevant atom-code. *JJ* coupling values are predicted for designated pairs of atoms and presented with equivalent statistical information.

Automatic Assignment of Chemical Shifts. A modification of the algorithm originally described for CSEARCH[3] is used to make automatic chemical shift assignments. The major improvement is the possibility of assignment of incomplete spectra where CSEARCH requires equal numbers of chemical shifts and atoms. Multiplicities arising from homo- and heteronuclear couplings are taken into consideration, and—a further improvement on the original—three assignment algorithms, *hard*, *soft*, and *force*, are available.

Hard uses the original (CSEARCH) algorithm to exclude false peak assignments, setting the minimum and maximum allowable chemical shifts by an arbitrarily-defined function (i.e., it has no physical or mathematical basis). *Soft* is in principle similar but bases the allowable chemical shift range

Table 3 *Hard*, *Soft*, and *Forced* Assignments: A Comparison of Three Variants of an Algorithm To Automatically Assign Chemical Shifts Demonstrated Against the Contents of the ^{19}F-File

			assignations		
method	structures	peaks	correct	incorrect	ambiguous
hard	20 760	114 458	99 054	940	14 464
soft	20 760	114 458	99 276	1032	14 150
force	20 760	114 458	109 028	5423	7

on the mean as opposed to the extreme values found in the database which reduces its sensitivity to erroneous assignations. Some peak assignments may remain ambiguous using both the hard and soft methods. Unique assignment of (nearly) all peaks can be *forced* by making assignments one-at-a-time and progressively increasing the tolerated deviation from values found in the database. This procedure yields the best statistics but can also generate erroneous—in particular inverted—assignations. In Table 3 a comparison of the results achieved when assignments of the structures in the ^{19}F-NMR file are calculated against the file for the three methods.

NMR Spectrum Simulation. Zero-order NMR spectra are displayed automatically on browsing through the database. If *JJ* coupling constants are defined, it is possible to calculate an NMR spectrum by diagonalization of the spin Hamilton matrix[13] which is constructed automatically, the *JJ* coupling constants being multiplied on all magnetically-equivalent nuclei. The Hamiltonian is divided into uncoupled blocks, the maximal size of each block being 64 (six coupled spins) in the current program version. Spectra are displayed as Lorentzian line shapes at the appropriate chemical shifts and intensities.

Substructure Search. Initially the atoms and bonds of the query fragment are sorted in the same way they are in structures, and then a brutto formula search (which is most efficient if the query fragment contains non-C atoms) is executed to eliminate irrelevant structures. Thereafter the atoms and bonds of the query fragment are matched sequentially to successive structures read from the database. The results of each attempted match are recorded in a pair of Boolean matrices with the dimensions *number of query atoms (bonds)* × *number of structure atoms (bonds)* which are populated with TRUE elements for a match and FALSE ones otherwise. Reference is made to the atoms matrix when populating the bonds one to ensure that TRUE bonds in fact join atoms of the correct type. If the matrices contain at least one TRUE element for each atom and bond of the query substructure a back-tracking algorithm,[14] starting with a maximally-coordinated non-C atom if one is present, is used to determine if the query maps to the structure.

This is a critical function in terms of its performance: although the time required rises with the number of nodes (atoms and bonds) in the substructure (exponential order), the time required to manipulate the Boolean matrices is proportional to their size (polynomial order), and this is a considerable advantage when retrieving elements from them when back-tracking to verify a possible match.

Chemical Shift Histogram for a Defined Atom. This NMR-specific search is implemented in the NIH/EPA Chemical Information System.[15] A query substructure is defined, and the atom of interest tagged. The search returns a histogram of the chemical shifts found, the mean, and standard deviation and enables the user to browse through the relevant structures.

***JJ* Coupling Constant Search.** This major search for *JJ* coupling constants is executed by defining a query substructure with a pair of tagged atoms to find all records containing *JJ* coupling constants for the fragment. Suppose one has an interest in ^2J(C,P) in vinylphosphonates: it is necessary to define vinylphosphonous acid as the query substructure and tag two atoms, phosphorus, and the vicinal carbon. The search will enable access to all records containing vinylphosphonate fragments with a defined value of ^2J(C,P).

The substructure search for *JJ* coupling constants differs from the substructure search based on subgraph isomorphism described above. The order of the *JJ* coupling constant must be conserved during the substructure search so result structures may not contain bonds which are absent in the query. For example, decalin is a valid result to a substructure query based on cyclodecane but invalid as a result of a search for *JJ* coupling constants because the bridging bond may change the order of the *JJ* coupling constant.

Results. Results, besides being viewed on screen, may be copied to an ASCII file of textural information on the records or a histogram of chemical shifts or printed.

NMR SPECTRUM LIBRARY

In the last decade several substantial collections of NMR spectra have been collated and are now commercially available as online and in-house products.[16-18]

They all contain spectra and structural information but surprisingly little data on coupling constants despite the enormous importance of this information for structure elucidation. They also betray a strong "Western" bias and do not exploit the "Eastern" literature which is hardly less extensive and indeed a good deal more complete for ^{19}F and ^{31}P. We have abstracted 30 journals (six English, the others Russian language) published in the countries of Eastern Europe and the former Soviet Union covering over 95% of publications in NMR spectroscopy in the period 1972−1992 and estimate that we abstracted not less than 97% of the relevant data in this literature. Excepting the Chemical Abstracts Registry number, which is not used in this literature, the data abstracted are shown in Figure 1. As an indication of the degree of chemical shift assignation we have constructed an assignment index (AI) for each record:

$$\mathrm{AI} = (N_{\mathrm{exact}} + 0.5N_{\mathrm{ambiguous}})/N_{\mathrm{total}}$$

where N_{total} is the total number of atoms of a given type in a molecule; N_{exact} is the number of exact chemical shift assignations; and $N_{\mathrm{ambiguous}}$ is the number of ambiguous chemical shift assignments.

Accordingly, a record has an index value of 1 if all the atoms of interest have defined chemical shift values and of 0.5 if all chemical shifts are presented but without assignment. AI may be <0.5 in cases where the list of chemical shift values is incomplete. From Table 4 it will be seen that by this criterion chemical shift assignation is extensive for all nuclei except ^{13}C, for which a histogram is presented in Figure 5.

Further investigation of the ^{13}C and ^{31}P records reveals that 79.6 and 83.2%, respectively, have connectivity matrices which are unique within the database. The repetitions are due to

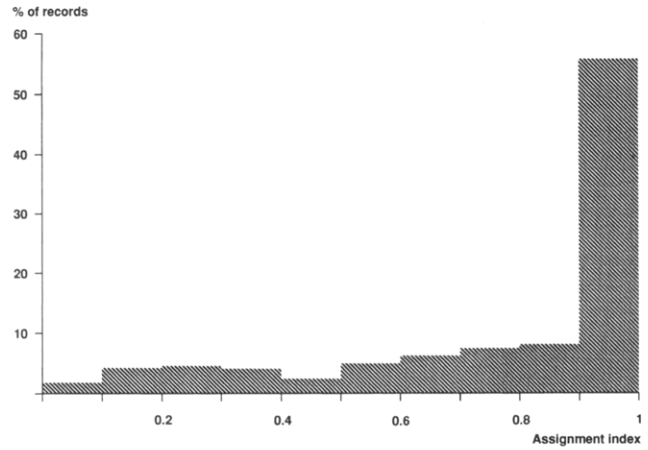

Figure 5. Assignment index distribution for the ^{13}C file.

Table 4. Database Content and Assignment Index (AI—See Text for Explanation)

nucleus	no. of records	no. of records with AI = 1	% records with AI = 1
^{11}B	1 374	1 094	79.6
^{13}C	36 184	18 344	50.7
^{15}N	1 617	1 365	84.4
^{17}O	748	663	88.6
^{19}F	23 800	20 613	86.6
^{29}Si	2 325	2 247	96.6
^{31}P	35 020	34 453	98.4
^{33}S	137	128	93.4

(1) structural isomers: 9.68% of ^{13}C and 3.44% of ^{31}P data

(2) different solvents: 3.71% of ^{13}C and 2.14% of ^{31}P data

(3) different temperatures: 0.42% of ^{13}C and 0.14% of ^{31}P data

(4) different authors: 0.79% of ^{13}C and 2.17% of ^{31}P data

(5) multiple publications: 5.63% of ^{13}C and 8.75% of ^{31}P data

CONCLUSION

WinDat is a powerful tool for abstraction and retrieval of NMR data, prediction, and assignment of chemical shifts and *JJ* coupling constants. Plausibility checks executed automatically during data input reduce typing and other errors in the database to an extremely low level and the software has been used to abstract the eastern European and Russian literature to create a unique database.

EXPERIMENTAL SECTION

WinDat comprises about 40 000 lines of code and is written in Borland Pascal. The program runs on an IBM-PC under microsoft Windows 3.1 and 8 Mbyte RAM and a VGA (or better) monitor. It requires 15 Mbyte hard disk for auxiliary files and 4.4 Mbyte per 10 000 database records (59 Mbyte for the full database of 101 205 spectra). Using a 50 MHz 80486DX2 PC, substructure (i.e., the slowest) searches run at 200 records/s for all-carbon substructures and 500 records/s if a heteroatom is present. In the worst case, a search through the 36 184 records of the ^{13}C-database will complete in under 3 min. Exact structure, substring, and numerical data searches are at least 5 times faster.

ACKNOWLEDGMENT

We express our gratitude to the Turpion-Moscow Co. for supplying the software used to write the program.

REFERENCES AND NOTES

(1) Bremser, W.; Fachinger, W. Multidimensional spectroscopy. *Magn. Reson. Chem.* **1985**, *23*, 1056−1076.
(2) Barth, A.; SpecInfo: An Integrated Spectroscopic Information System. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 52−58.
(3) 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.
(4) Sasaki, S.-I.; Kudo, Y. Structure Elucidation System Using Structural Information from Multisources: CHEMICS. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 252−257.
(5) Elyashberg, M. E.; Serov, V. V.; Martirosyan, E. R.; Zlatina, L. A.; Karasev, Yu. Z.; Koldashev, V. N.; Yampolskiy, Yu. Yu. An Expert System for Molecular Structure Elucidation Based on Spectral Data. *J. Molec. Structure (Theochem)* **1991**, *230*, 191−203.
(6) Product information may be obtained from Chemical Concepts GmbH, P.O. Box 10 02 02, D-69442, Weinheim, FRG, Tel. +49 6201 606433, Fax +49 6201 606430. The information in this publication may not be construed as a commitment ("zugesicherte Eigenschaft") by Chemical Concepts GmbH.
(7) Glebko, L. I.; Berezhevskaya, L. I.; Ul'kina, Zh. I.; Strigina, L. I.; Zinova, S. A. Optimization of getting of 25R-spirost-5en-3β,17α-diol (pennogenin) from flowers. *Khim. Prirod. Soedin. (Russian)* **1987**, 115−119.
(8) Hanessian, S.; Franco, J.; Gagnon, G.; Laramee, D.; Larouche, B. Computer-Assisted and Perception of Stereochemical Futures in Organic Molecules Using the CHIRON Program. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 413−425.
(9) Standard Practice for Data Presentation Relating to High-Resolution Nuclear Magnetic Resonance (NMR) Spectroscopy. *Reprint from the Annual Book of ASTM Standards*; 1916 Race Street, Philadelphia, PA 19103.
(10) Gasteiger, J.; Hendriks, B. M. P.; Hoever, P.; Jochum, C.; Somberg, H. JCAMP-CS: A standard exchange format for chemical structure information in computer-readable form. *Appl. Spectrosc.* **1991**, *45*, 4−11.
(11) Davies, A. N.; Lampen, P. JCAMP-DX for NMR. *Appl. Spectrosc.* **1993**, *47*, 1093−1099.
(12) Bremser, W. HOSE—a novel substructure code. *Anal. Chim. Acta* **1978**, *103*, 355−365.
(13) Gunther, H. *NMR Spectroscopy. An. Introduction.* John Wiley & Sons: Chichester, 1980; 477.
(14) Barnard, J. M. Substructure Searching Methods: Old and New. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532−538.
(15) Heller, S. R. The Chemical Information System and Spectral Databases. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 224−231.
(16) Heller, S. R. Computerized spectroscopy databases. *Chemistry International.* **1991**, *13*, 235−238.
(17) Warr, W. A.; Suhr, C. Chemical Information Management. VCH Verlagsgesellschaft: Weinheim and New York, 1992; p 261.
(18) Warr, W. A. Computer-assisted structure elucidation. *Anal. Chem.* **1993**, *65*, 1045.

CI940346R