Available online at www.sciencedirect.com
ScienceDirect
Journal of Magnetic Resonance 188 (2007) 97–110
JMR Journal of Magnetic Resonance
www.elsevier.com/locate/jmr

Identification of organic molecules from a structure database using proton and carbon NMR analysis results

Reinhard Dunkel *, Xinzi Wu

ScienceSoft LLC, 9934 Pinehurst Drive, Sandy, UT 84092, USA

Received 15 January 2007; revised 4 June 2007
Available online 30 June 2007

Abstract

A compound is identified by matching its proton and/or carbon NMR spectra to NIH PubChem molecular structures. The matching process involves analyzing 1D proton, 1D carbon, DEPT, and/or HSQC spectra, and comparing the number of NMR resonances, detected proton and carbon shifts, likely number of methyl- and methoxy-groups, and an optionally specified molecular formula to predicted proton and carbon shifts of PubChem structures. A structure verification module rates the consistency between experimental spectral analysis results and a proposed structure (not limited to PubChem structures) and assigns observed shifts to the proposed structure. The spectral analysis, structure identification, and structure verification are largely automated in a software package and can be performed in minutes.

Keywords: Structure identification; Dereplication; Structure verification; Structure elucidation; PubChem

1. Introduction

Nuclear Magnetic Resonance (NMR) is an important technique to elucidate unknown organic molecules. But its sensitivity is limited and many elements lack an NMR observable isotope. So a full structure elucidation relies on several spectroscopic techniques and is normally performed by an experienced spectroscopist. Automated approaches for structure generation [1] and elucidation [2] start to appear, but their success rates remain limited.

Given most published small organic molecular structures, the structure elucidation of an unknown molecule is replaced by searching for best matching structures from this collection. Organic molecules consist mainly of proton and carbon atoms. An observed chemical shift reflects the chemical environment an atom experiences in the sample. Nearby atoms of any type influence its shift. So small organic molecules can be identified by matching the observed proton and carbon shifts to the predicted shifts of candidate structures. If the molecule is contained in the candidate structure database, it is likely identified as the best possible match. Should it not be contained, identified best matches tend to be similar in molecular size and functionality. This structure identification approach can be performed with limited information about the unknown and can be reliably automated. It is fast to determine whether an unknown compound is a published structure and full structure elucidations can be limited to a few challenging cases.

We implemented the structure identification process in a software package, NMRanalyst. The data analysis module of the NMRanalyst software extracts numerical description information (including chemical shifts) from NMR data. It transforms, baseline and phase corrects, and analyzes NMR data with minimal user interaction. Its analysis sensitivity often exceeds the visual inspection by an experienced spectroscopist [3]. The NMRanalyst FindIt module contains over 8 million National Institutes of Health (NIH) PubChem structures and the predicted proton and carbon shifts for these structures. To identify a compound

* Corresponding author. Fax: +1 801 816 0163.
E-mail address: dunkel@sciencesoft.net (R. Dunkel).

1090-7807/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2007.06.007

from the FindIt structure database, the analyzed and extracted numerical information of the compound is compared with the predicted shifts of each candidate structure and the consistency is rated by the NMRanalyst VerifyIt module. FindIt lists the best matching structures with the top VerifyIt ratings. Fig. 1 illustrates the interactions among these NMRanalyst modules.

The structure identification by NMRanalyst is evaluated using 179 organic molecules. The obtained FindIt placement for the correct structure is determined for five common input combinations. As the acquisition of a one-dimensional (1D) carbon spectrum is about 5700 times less sensitive than that of a corresponding 1D proton spectrum, more sensitive carbon acquisition schemes are evaluated.

2. Methods

2.1. Evaluation compounds

One hundred and seventy-nine compounds are used for this evaluation. They range from structures of a few atoms to larger pharmaceuticals such as fexofenadine, lasalocid, and taxol. For half of them (81 compounds), NMR datasets are analyzed by NMRanalyst. The spectra are from samples containing one major compound. For the other half (98 compounds), shifts and proton integrals are obtained from web sites, such as WebSpectra (http://www.chem.ucla.edu/~webspectra) and SDBS (http://www.aist.go.jp/RIODB/SDBS/menu-e.html), books, and other printed sources. NMRanalyst converts such information to resonance descriptions for use in this evaluation.

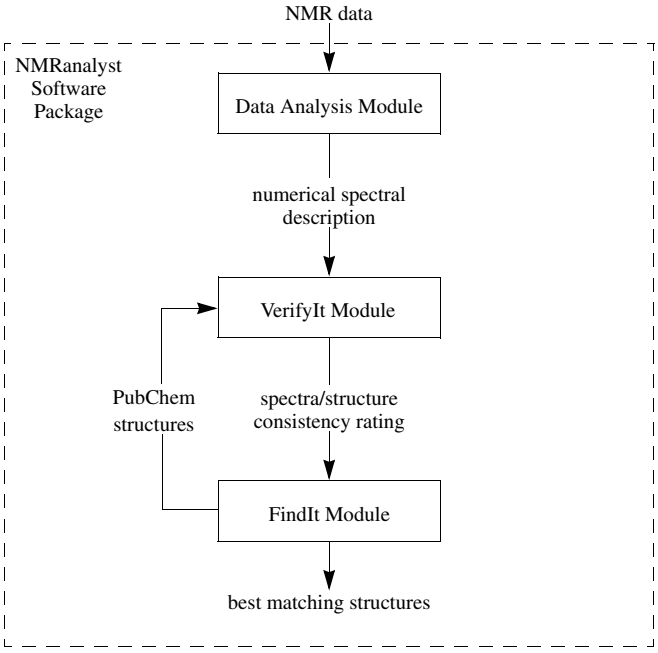

Fig. 1. Interactions among the NMRanalyst software modules for structure identification.

Sixty-seven evaluation datasets originate from WebSpectra and were potentially unreferenced. Carbon spectra are referenced based on locking solvent resonances. For a proton spectrum, the tetramethylsilane (TMS) resonance is used for referencing, when detected. Otherwise, a noticeable HDO or residual undeuterated solvent resonance is used. If neither approach applies, the average difference between observed and predicted proton shifts for the specified structure is used (see Section 2.3 for details).

2.2. Analysis of NMR datasets

We developed the data analysis module of the NMRanalyst software to analyze the NMR data. NMR datasets are acquired in the time domain and are transformed to the frequency domain using a Fourier Transform. Spectral baseline and phase are automatically corrected. For underdigitized or low signal-to-noise spectra, this automated correction is more reliable than a visual one [4]. The results of the data analysis step are the NMRanalyst generated numerical spectral descriptions.

The analysis of a 1D proton or 1D carbon spectrum starts with an initial peak-picking step. Identified possible resonances are fitted in the complex valued spectrum using a Lorentzian shape model, representing the spin relaxation, convoluted by a sinc function resulting from the finite acquisition time. Clusters of overlapping resonances are fitted simultaneously, so the numerical description of a 1D spectrum reflects the mutual resonance overlap. The shift value is the resonance frequency determined through the best-fit of this resonance model to the experimental data. Its accuracy is better than the spectral resolution of the analyzed spectrum [5].

Fig. 2 shows the taxol structure with its 1D proton (top) and 1D carbon (bottom) spectra. The baseline and phase corrected experimental spectrum is drawn in yellow (shown as gray in the black-and-white print-out). Each resonance is modeled by its resonance frequency, absolute value integral, relaxation time, phase, and the acquisition time. As a visual representation of the determined numerical signal descriptions, a simulated spectrum is calculated as the sum of the best-fit individual spectral resonances. Intuitively, a simulated spectrum is a re-creation (or simulation) of the original experimental spectrum from the numerical description of each resonance in the experimental spectrum. The simulated spectrum is drawn on top of the experimental spectrum in black as shown in Fig. 2. Ideally, no experimental resonances (except for noise) should remain visible in such a display, indicating that the NMRanalyst generated numerical descriptions adequately model the experimental resonances.

For some evaluation compounds, protonated carbon shifts are determined from a Distortionless Enhancement by Polarization Transfer (DEPT-135) or Heteronuclear Single Quantum Coherence (HSQC) spectrum. Fig. 3 shows the gibberellic acid structure and its 1D proton and HSQC spectra. The 1D proton spectrum is drawn at

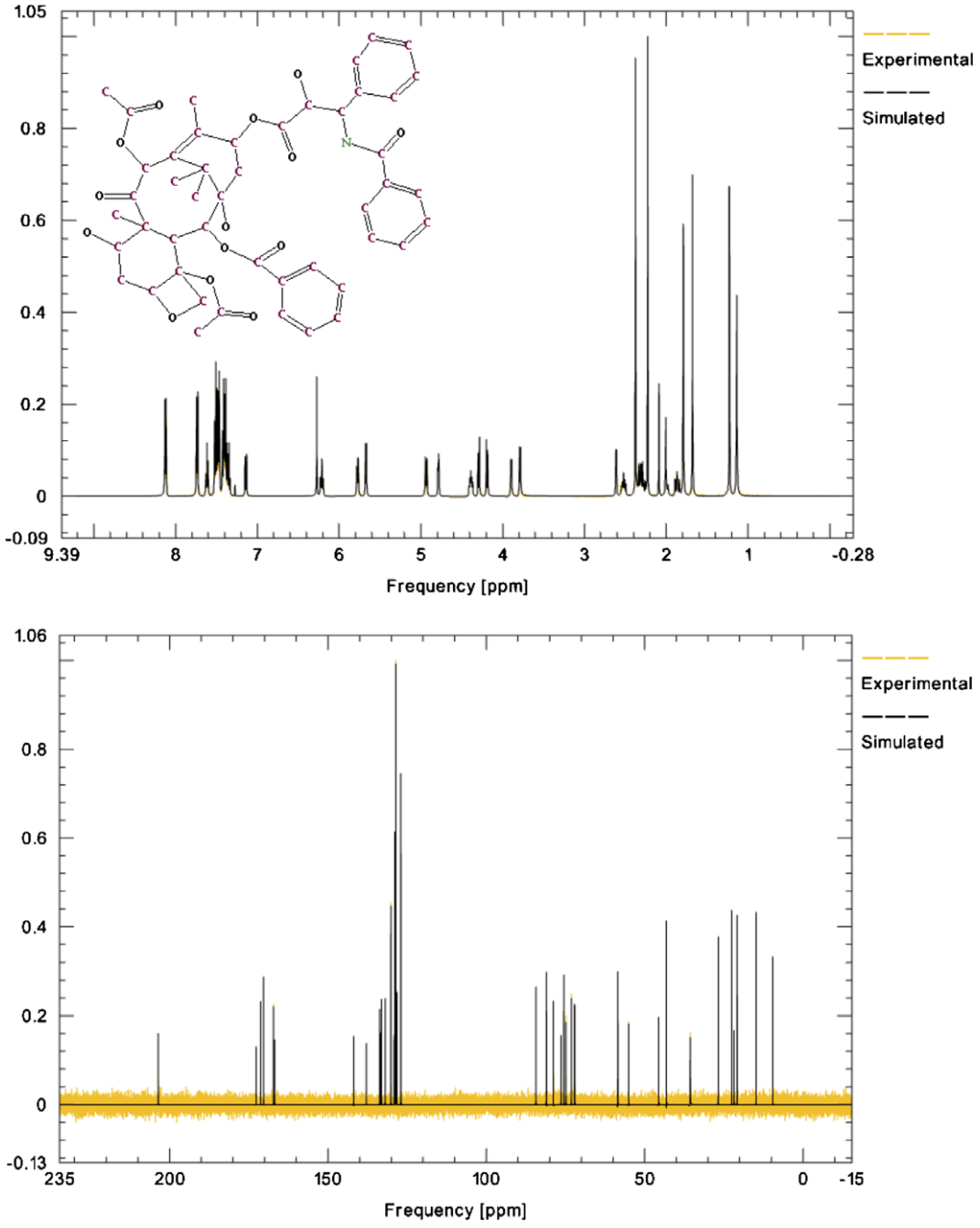

Fig. 2. Taxol evaluation compound with molecular structure and Varian INOVA 500 MHz acquired 1D proton (top) and 1D carbon (bottom) spectra of a 8.3 mg sample.

the top along the F2 dimension of the HSQC spectrum. The rectangles in the HSQC spectrum indicate fitting areas, in which NMRanalyst detects a correlation. While HSQC is a 2D spectrum, its acquisition tends to be several times faster than that for a corresponding 1D carbon spectrum. A good F1 resolution is time consuming to acquire in a 2D spectrum. NMRanalyst improves the frequency determination by modeling the shape of each resonance and fitting it simultaneously in both spectral dimensions and all four spectral phase components. Without strong resonance overlap, carbon shifts are determined with an accuracy better than a fifth of the acquired F1 spectral resolution.

A common NMR practice is to use Linear Prediction resolution enhancement for under-digitized spectra [6]. But Linear Prediction modifies the apparent acquisition time of individual resonances and introduces additional spectral point correlations. It should not be used in combination with the NMRanalyst modeling of spectral correlations.

2.3. VerifyIt consistency rating and shift assignments

The NMRanalyst VerifyIt module explains the consistency between the NMR data and a proposed structure.

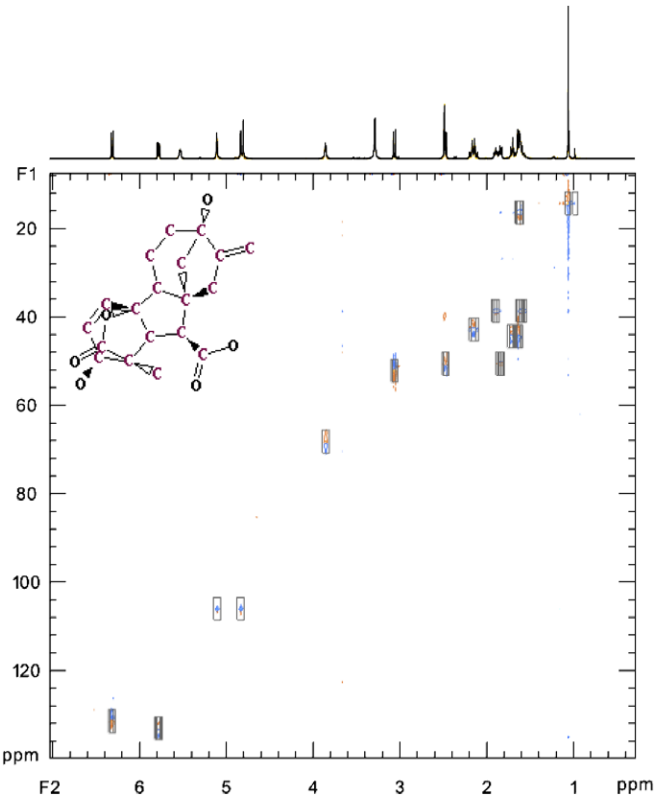

Fig. 3. Gibberellic acid evaluation compound with molecular structure and Varian INOVA 500 MHz acquired 1D proton and HSQC spectra of a 7 mg/ml sample. Rectangles enclose HSQC resonances from which the protonated carbon frequencies are derived.

VerifyIt rates the consistency based on (1) the observed vs. expected numbers of proton and carbon resonances, (2) the observed vs. expected numbers of methyl- and methoxy-groups, (3) the observed vs. predicted proton and carbon shifts, and (4) the spectral purity rating.

To compare lists of chemical shifts with different lengths, individual entries from the shorter list are duplicated (assuming different nuclei have the same shift due to accidental degeneracy) until the length of both lists agree and the average deviation between the two lists obtains the minimal value. VerifyIt uses the average shift deviation between the observed and predicted shifts for a possible experimental spectrum referencing correction.

Two criteria are used to rate the consistency between the observed and predicted chemical shifts: the maximum shift deviation and the average shift deviation. Due to shift prediction imperfections, the average deviation is given a stronger weighting than the maximum deviation in the consistency rating. The proton shift rating is compromised by the limitation that only the carbon-bonded protons in a candidate structure are predicted, while a proton spectrum may contain heteroatom-bonded proton resonances.

The spectral purity rating is determined from the 1D proton spectrum, as its integrals are fairly quantitative. The proton spectrum is analyzed and residual solvent, water, and TMS resonances are removed. The remaining resonances are separated into clusters of overlapping resonances. The numbers of carbon-bonded and heteroatom-bonded protons are known from the candidate structure. We expect carbon-bonded protons to be observed, while heteroatom-bonded protons may or may not be detected.

The calculation of a purity rating starts with the estimation of the average integral of a single proton. If the candidate structure only contains carbon-bonded protons, this average proton integral is the total spectral integral divided by the number of protons in the structure. If the proposed structure also contains heteroatom-bonded protons, the average proton integral lies between the total spectral integral divided by the total number of protons and the total spectral integral divided by the number of carbon-bonded protons. The integral for each resonance cluster is normalized by the estimated average proton integral. If the purity rating is perfect, an integer would result from the normalization of each cluster integral by this average integral, representing the number of protons in each cluster. In case that a decimal number results, the fractional part of the number (or one minus the fractional part, whichever is less) is taken as a measurement of impurity and is added up for all the clusters. The greater the sum of the fractional parts, the lower the resulting purity rating. The final estimate for the average proton integral for a structure containing heteroatom-bonded protons is the value within the possible range resulting in the highest purity rating. This approach for purity rating may not be ideal, but it is reliable and fast to determine.

The relative weighting among the various factors for deriving a consistency rating varies for proton, protonated carbon (DEPT-135 or HSQC), and carbon analysis results. The carbon shift prediction is more reliable than the proton shift prediction. Furthermore, only carbon-bonded (but not heteroatom-bonded) proton shifts are predicted for a candidate structure (see Section 2.5 for details). Poor F1 resolution of an HSQC spectrum limits the number of detected resonances, whereas a 1D carbon spectrum has a significantly higher resolution. The relative weighting is also affected by whether a candidate structure contains fluorine or phosphorus atoms (see Section 2.7). The evaluation datasets were used to optimize the weighting parameters. However, a parameter optimized for the idiosyncrasy of one dataset is unlikely to result in reasonable ratings for other datasets. As we continue to collect more datasets, these parameters will be further optimized.

Besides generating a consistency rating, VerifyIt also determines whether a proposed structure explains the observed NMR data better than the millions of structures in the FindIt structure database. VerifyIt reports a placement for the proposed structure relative to all the FindIt structures with regard to how well the structure matches the NMR data. If the candidate structure obtains the place one, which means that it agrees with the specified NMR data better than any of the 8 million FindIt structures, it is a good indication that the candidate structure is likely the correct one. Verification approaches not involving this

comparison with competing structures have been published before [7,8].

VerifyIt can assign detected chemical shifts to a proposed structure. To carry out the assignment, the carbon and carbon-bonded proton shifts are predicted for the specified structure (see Section 2.5). Each predicted shift is then replaced by the closest matching observed shift. As the heteroatom-bonded proton shifts are not predicted, any detected heteroatom-bonded proton shifts are unlikely to match predicted carbon-bonded proton shifts and are simply ignored. Proton–proton couplings complicate the interpretation of multiplets in a proton spectrum. The median frequency of the proton resonance cluster is taken as the observed proton shift value and is assigned based on the closest match to a predicted proton shift. Fig. 4 displays the assigned carbon-bonded proton (top) and carbon (bottom) shifts for the taxol structure from the analysis of the 1D proton and carbon spectra in Fig. 2.

2.4. PubChem structure selection and FindIt structure database

NIH introduced PubChem in September 2004 to provide information about small molecules and their biological activities (http://nihroadmap.nih.gov). In April 2007, over 10 million PubChem structures were downloaded from ftp://ftp.ncbi.nlm.nih.gov/pubchem.

Table 1 shows the elements, for which we found bonded proton and carbon shift information [9,10], in bold. Structures consisting only of these element types, having one through 100 carbons and no more than 256 protons, and representing a single structure or a main structure with additional fragments of no more than one non-proton atom, possibly charged, are retained. For example, structures with an additionally specified HCl or an atom anion or cation are included in the FindIt structure database.

PubChem structures specify bonded hydrogens to unambiguously represent each molecule. But most element types only have one stable number of valences. Unspecified valences are assumed to represent a bond to an implied hydrogen atom. For example, a single specified carbon is interpreted as methane, CH_4. A single nitrogen is interpreted as ammonia, NH_3. But CHg_4 would be misinterpreted as CH_4Hg_4 using this simplified structure representation. PubChem structures are specified observing the Octet Rule. Among the supported atom types in Table 1, only the elements with multiple valences in the singly negative, uncharged, or singly positive state have the bonded protons retained in the FindIt structures to avoid misinterpretations. These elements are Si, P, S, Cl, Ti, Cr, Fe, Cu, Ge, As, Se, Br, Ru, Rh, In, Sn, Sb, Te, I, Hg, Tl, Pb, and Bi (S.E. Stein, S.R. Heller, D.V. Tchekhovskoi, The IUPAC Chemical Identifier—Technical Manual, Appendix 1, InChI Standard Valences, http://www.iupac.org/inchi).

The International Union of Pure and Applied Chemistry (IUPAC) introduced the International Chemical Identifier

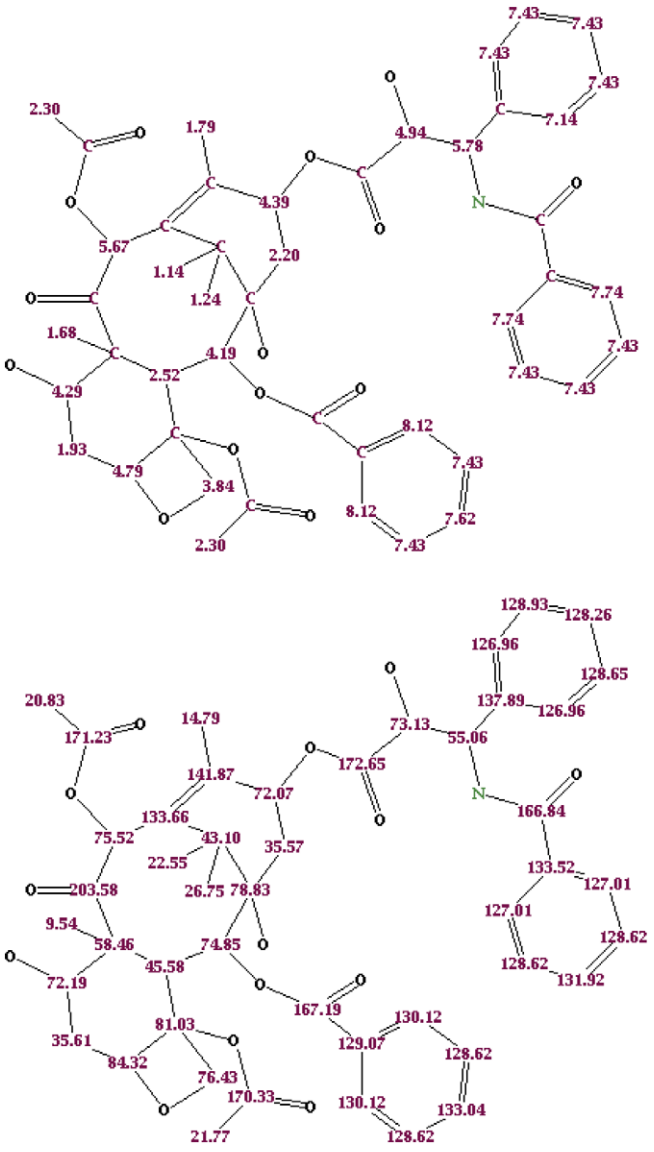

Fig. 4. Taxol structure with the VerifyIt assigned proton (top) and carbon (bottom) shifts in ppm determined from the 1D proton and carbon spectra in Fig. 2.

(InChI) to provide a unique text representation for a molecular structure. After removing structures with unsupported atom types and eliminating radical, isotope, and unsupported stereo chemistry specifications, we used the InChI software (downloaded from http://www.iupac.org/inchi) to create a canonical text string representation for the remaining structures. When several structures result in the same InChI string, only the one with the smallest PubChem Compound ID (CID) is retained. From the downloaded PubChem structures, over 8 million unique structures remain.

To put this number of FindIt structures into perspective, a nearly unlimited number of small organic molecules could be synthesized. The Chemical Abstract Service Registry assigns CAS numbers to chemical compounds mentioned in scientific publications and patents. For the

Table 1
Elements contained in FindIt structures are shown in this periodic table of the elements in bold

H																He
Li	Be										**B**	**C**	**N**	**O**	**F**	Ne
Na	**Mg**										**Al**	**Si**	**P**	**S**	**Cl**	Ar
K	**Ca**	Sc	**Ti**	V	**Cr**	Mn	**Fe**	Co	Ni	**Cu**	**Zn**	**Ga**	**Ge**	**As**	**Se**	**Br**	Kr
Rb	**Sr**	Y	Zr	Nb	Mo	Tc	**Ru**	**Rh**	Pd	**Ag**	**Cd**	**In**	**Sn**	**Sb**	**Te**	**I**	Xe
Cs	**Ba**		Hf	Ta	W	Re	Os	Ir	Pt	Au	**Hg**	**Tl**	**Pb**	**Bi**	Po	At	Rn
Fr	Ra																

described selection criteria, the Registry contains around 9 million CAS numbers. The 8 million FindIt structures are equivalent to most published small organic molecule structures. A user can add further unique structures to the FindIt database.

2.5. Proton and carbon shift prediction

The carbon and carbon-bonded proton shifts are predicted for all the FindIt structures. The heteroatom (non-carbon) bonded protons are often labile and their NMR resonances tend to be broad and may not be distinguishable from spectral baseline distortions. FindIt does not predict or use them to determine structural matches.

NMRanalyst shift prediction uses additivity rules and Hierarchical Organization of Spherical Environments (HOSE) codes with assigned shifts. Pretsch published a set of additivity rules for predicting carbon and proton shifts [10–15]. These rules are implemented in the NMRanalyst software. Further refinements are based on Bremser HOSE code representations of assigned shifts [16,17]. Assigned carbon and proton shifts are gathered from on-line databases, such as NMRShiftDB [18] (http://www.nmrshiftdb.org), the Japanese National Institute of Advanced Industrial Science and Technology (AIST) Spectral Database for Organic Compounds (http://www.aist.go.jp/RIODB/SDBS/menu-e.html), NMRDBTech (http://www.las.jp/products/chnmrnp/CH-NMR-NP/index.html), publications, and books. For protons, only carbon-bonded proton shifts are used and are assigned to the bonded carbon. When two methylene proton shifts differ, their average value is assigned to the bonded carbon.

Some of the evaluation compounds used in this study have published assigned shifts. To obtain structure identification results representative for unknown compounds, assigned shifts for any evaluation compounds were removed and did not contribute to the shift prediction.

A six-sphere HOSE code is created for each assigned shift to represent the corresponding chemical environment. The higher the sphere, the more distant the sphere is from the center atom with the assigned shift. Shifts with the identical HOSE code are aggregated to derive the median shift and shift range. Over 900,000 unique six-sphere HOSE codes with assigned carbon shifts and over 45,000 with assigned proton shifts are derived. Next, the highest sphere (i.e., the sixth sphere) is truncated to create a five-sphere HOSE code for the associated shift value. Again, identical five-sphere HOSE codes are aggregated. This sphere truncation process continues until one-sphere HOSE codes are created and the associated assigned shifts are aggregated. (The Bremser notation for carbonyl groups is used in one-sphere HOSE codes.)

To predict a carbon or proton shift, the six-sphere HOSE code is generated for the atom of interest. If an exact match of the six-sphere HOSE code can be found with assigned shifts, the median value of the assigned shifts is taken as the predicted shift. If no exact match can be found, the six-sphere HOSE code is truncated to five-sphere and the search for the exact match is repeated for the five-sphere HOSE code. The higher the matching sphere, the smaller the associated shift range, and hence the more accurate the predicted shift value. When less than three HOSE spheres match, the additivity rule prediction is applied. If needed additivity rules are missing, two- or even one-sphere HOSE code predictions are attempted. While lower sphere HOSE predictions nearly always succeed, their associated shift ranges tend to be large, and the resulting predicted shift may be inaccurate. This prediction algorithm applies to both carbon and carbon-bond proton shift prediction.

Although stereo chemical information is available for the FindIt structures, this information is not currently used for shift prediction. The predicted shift values are stored together with the corresponding structures in the FindIt database.

2.6. FindIt best structure matches

After the NMR spectra of an unknown compound are analyzed, the FindIt module of NMRanalyst identifies the structures best matching the experimental data. The predicted carbon and proton shifts of each FindIt structure are rated in consistency with the observed NMR data. For the taxol dataset (see Fig. 2), FindIt reports the top 10 structure matches, listing the obtained placement, determined structure rating, and PubChem CID in parentheses as shown in Table 2.

Fig. 5 shows the corresponding structure display. The correct taxol structure is identified at place one and displayed at the top left corner in Fig. 5. Further information about the matched structures and their biological activities can be obtained from http://pubchem.ncbi.nlm.nih.gov/ search by specifying the listed CID numbers.

Table 2
FindIt program output for structures best matching the 1D proton and carbon spectra of taxol (see Fig. 2)

Best 10 structures in decreasing rating (structure ID shown in parentheses):		
1: 0.957244 (4666)	2: 0.955890 (9940855)	3: 0.955832 (10328320)
4: 0.954903 (3426646)	5: 0.954752 (10260115)	6: 0.954258 (3694420)
7: 0.954149 (10350793)	8: 0.954056 (10418463)	9: 0.953968 (10033647)
10: 0.953893 (11332028)		

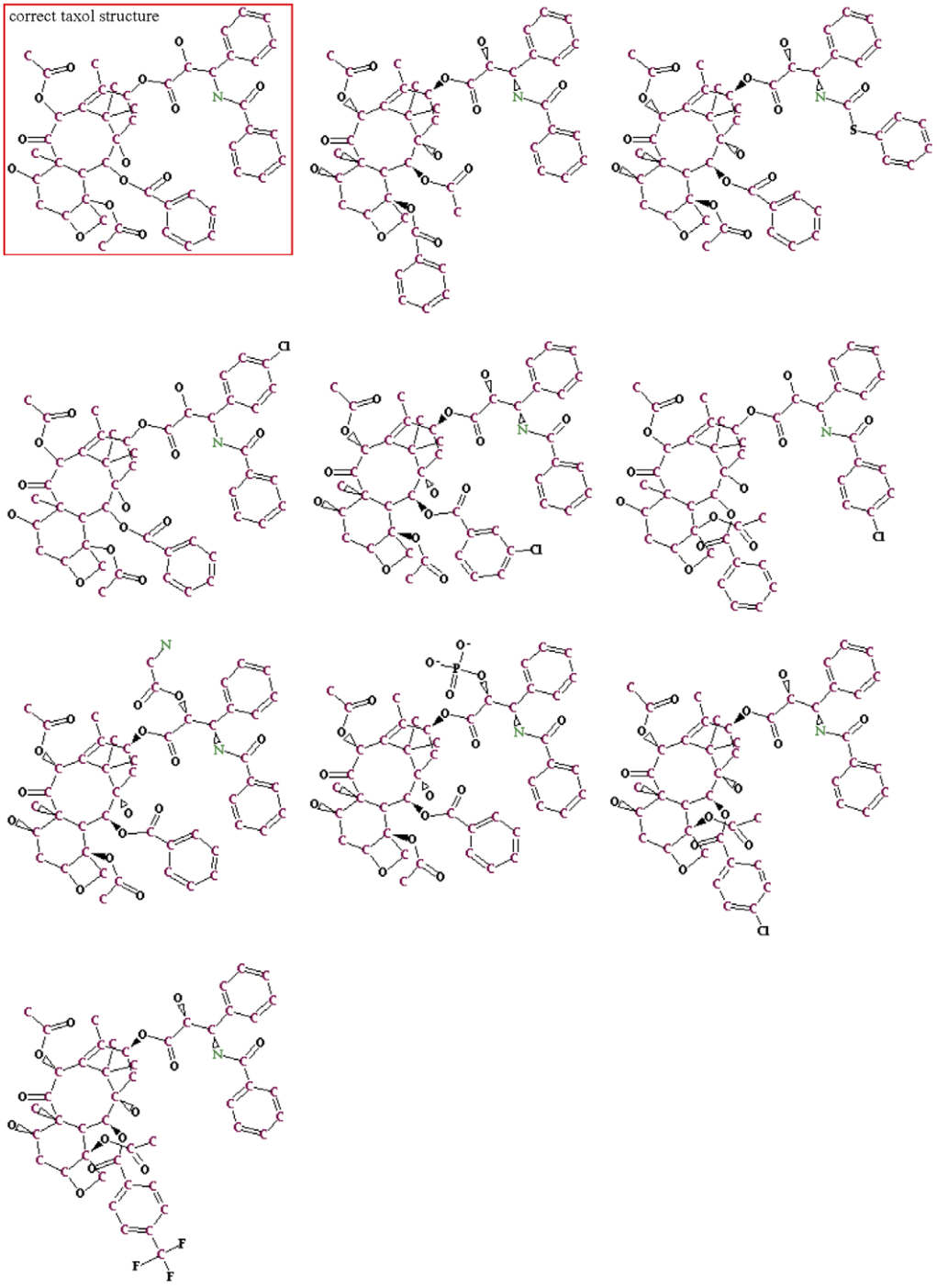

Fig. 5. Top 10 matching molecular structures identified from the 8 million FindIt structures based on the 1D proton and carbon spectra of taxol (see Fig. 2). The top left structure is the best match and is the correct taxol structure.

2.7. ^{19}F and ^{31}P couplings

Fluorine and phosphorus consist of the NMR active ^{19}F and ^{31}P isotopes, respectively. A triple resonance probe could be used to broadband decouple resulting splittings. As this hardware is uncommon, NMRanalyst provides two software solutions.

Known couplings can be removed from observed spectra. Fig. 6 displays the $C_7H_{14}FNO$ structure, and fluorine coupled (top) and decoupled (bottom) carbon spectra. The bottom spectrum is calculated by subtracting the detected fluorine coupled 82 and 58 ppm carbon resonances and adding decoupled resonances with an average frequency, relaxation time, phase, and the sum of the previously splitted resonance integrals back in the spectrum. Such a coupling removal is needed for VerifyIt to assign the detected shifts to a proposed structure.

A more general approach is to consider a resulting unknown number of resonances and couplings for structures containing fluorine or phosphor atoms. This approach is implemented in FindIt. Several of the evaluation compounds contain ^{19}F or ^{31}P nuclei. FindIt usually identifies the correct structure without removal of the couplings through experimental or software means.

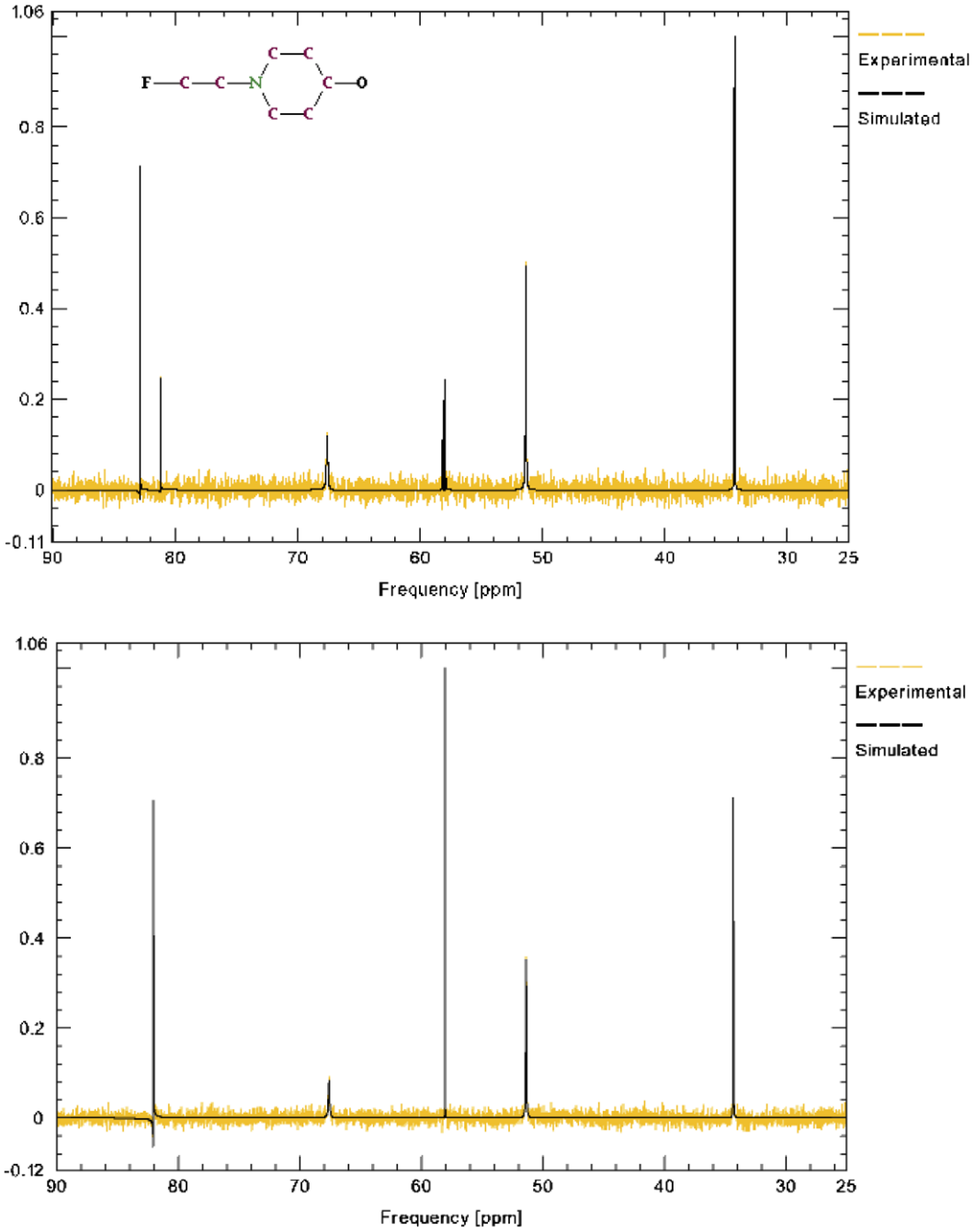

Fig. 6. $C_7H_{14}FNO$ evaluation compound with molecular structure and Bruker AMX 400 MHz acquired fluorine coupled (top) and NMRanalyst software fluorine decoupled (bottom) carbon spectra.

Table 3
FindIt rankings of evaluation compounds for five common input combinations

#	Dataset	PubChem CID	Structures same MF	^{1}H, MF	^{1}H, $^{13}C_H$	^{1}H, $^{13}C_H$, MF	^{1}H, ^{13}C	^{1}H, ^{13}C, MF
1.	1-Hexyne	12732	56	1			1	
2.	1-Indanone	6735	59	1	1		1	
3.	2-Bromoanisole	11358	23	1			2	1
4.	2-Butanone	6569	28	1			1	
5.	2-Ethyl-1-indanone	640236	293	4	522	1	1	
6.	2-Methylpentan-3-one	11265	158	1			1	
7.	2-Phenylethanol	6054	131	1			1	
8.	3-Bromoanisole	16971	23	1			7	1
9.	3-Bromocyclohexene	137057	25	1			13	1
10.	3-Nitrophenol	11137	32	6			2	1
11.	4-Bromoanisole	7730	23	1			31	1
12.	4-Penten-1-ol	13181	70	1			1	
13.	Acrolein	7847	8	1			1	
14.	Allyl alcohol	7858	10	1			3	1
15.	Anthranilic acid	227	122	1			39	1
16.	Benzophenone	3102	48	1			1	
17.	Benzopurpurin 4B	13817	5	2			2	2
18.	Benzothiazole	17428	20	1			1	
19.	Berberine	2353	49	1			1	
20.	Brucine	9649	1518	4	1		2	1
21.	Butyl acetate	31272	207	1			1	
22.	C_3H_7ClO	12313	7	1			1	
23.	C_3H_9NO	5126	16	2			3	2
24.	C_4H_6O	6570	26	1			1	
25.	$C_4H_7BrO_2$	76934	14	1			1	
26.	C_4H_7IO	5324487	19	1			1	
27.	C_4H_8O	69389	28	1			1	
28.	$C_4H_{10}O_2$	75103	24	1			1	
29.	C_5H_8O	8080	76	1			1	
30.	$C_5H_8O_2$	98451	136	5			31	3
31.	$C_5H_{14}OSi$	18013	5	1			1	
32.	C_6H_4BrI	11415	3	1			1	
33.	C_6H_6BrN	11562	17	1			2	1
34.	$C_6H_{10}O_2$	7838	283	1			1	
35.	$C_6H_{12}O$	5324489	158	1			1	
36.	C_7H_7Br	11560	6	3			243	3
37.	C_7H_7N	7502	17	1			1	
38.	$C_7H_8O_2$	9007	126	1			2	1
39.	$C_7H_{13}NO_5$	300859	40	2			3	1
40.	$C_7H_{14}FNO$	450035	6	1			726	1
41.	$C_7H_{16}O_4$	66019	28	1			1	
42.	$C_8H_5F_3O_2S$	U12	7	1			5	1
43.	$C_8H_8O_2$	8658	134	1			4	2
44.	$C_8H_8O_3$	10322	158	1			1	
45.	$C_8H_{10}Cl_2O_2$	641639	13	1			2	1
46.	$C_8H_{10}O_2$	7149	336	1			1	
47.	$C_8H_{12}O_2$	137957	599	2			3	1
48.	$C_8H_{12}O_2$_b	87859	599	1			1	
49.	$C_8H_{17}O_5P$	13345	26	1			1	
50.	C_9H_8BrNO	12793313	49	5			1	
51.	$C_9H_{10}O_2$	31217	288	3			1	
52.	$C_9H_{13}NO$	76652	361	1			1	
53.	$C_{10}H_7Br$	11372	5	1			3	1
54.	$C_{10}H_7BrO$	11615504	18	10			9	1
55.	$C_{10}H_8BrN$	2795554	31	1			25	1
56.	$C_{10}H_{10}O_2$	95942	368	1			1	
57.	$C_{10}H_{10}O_2$_b	15173	368	1			1	
58.	$C_{10}H_{10}O_2$_c	101335	368	27			3	1
59.	$C_{10}H_{11}NO_4$	24047	411	1	1		1	
60.	$C_{10}H_{12}O_2$	94247	600	1			4	1
61.	$C_{10}H_{12}O_3$	34656	598	1	1		1	
62.	$C_{10}H_{15}N$	7061	199	1			1	
63.	$C_{10}H_{16}O_2$	7561	1027	5			1	

(*continued on next page*)

Table 3 (*continued*)

#	Dataset	PubChem CID	Structures same MF	^{1}H, MF	^{1}H, $^{13}C_H$	^{1}H, $^{13}C_H$, MF	^{1}H, ^{13}C	^{1}H, ^{13}C, MF
64.	$C_{10}H_{18}O$	7392	584	9			1	
65.	$C_{11}H_{11}BrO_5$	U10	14	1			4	1
66.	$C_{11}H_{11}NO_2$	3744	554	1			1	
67.	$C_{11}H_{12}O_2$	260944	596	2			1	
68.	$C_{11}H_{12}O_3$	5323697	615	2	16	1	1	
69.	$C_{11}H_{16}N_2O_3$	5323896	372	21			3	1
70.	$C_{11}H_{17}F_3O_5S$	643576	2	1			2	1
71.	$C_{12}H_7N_3O_2$	5326134	35	1	744	1	17	1
72.	$C_{12}H_{12}N_2O_2S$	5323873	509	7			17	1
73.	$C_{12}H_{12}N_2O_3$	U3	637	9			5	2
74.	$C_{12}H_{12}O_2$	11788723	465	1	14	1	13	1
75.	$C_{12}H_{13}N_3O_2$	693059	634	2			740	2
76.	$C_{12}H_{14}N_2O$	85812	541	2			1	
77.	$C_{12}H_{15}NO_2$	38362	908	1	2	1	1	
78.	$C_{12}H_{23}NOSSi$	643605	2	1			1	
79.	$C_{13}H_{12}O$	10976167	125	1	745	1	18	1
80.	$C_{13}H_{14}O_2$	5323698	525	1	1		1	
81.	$C_{13}H_{17}NO_2$	11608231	997	11			1	
82.	$C_{13}H_{19}N_3O_2$	5323898	400	1			181	1
83.	$C_{14}H_{12}N_2O_2$	5291723	544	4			1	
84.	$C_{14}H_{14}$	7647	79	1			5	1
85.	$C_{14}H_{19}N_3O_3$	5323897	498	2			1	
86.	$C_{14}H_{21}N_3O_2$	U9	408	6			445	1
87.	$C_{14}H_{21}N_3O_3$	U6	348	32			1	
88.	$C_{15}H_{11}ClO_3$	U18	91	5	1		1	
89.	$C_{15}H_{20}O_2$	5326120	743	1			1	
90.	$C_{16}H_{11}ClO_2$	643579	59	23			2	1
91.	$C_{16}H_{12}N_2$	U14	143	3			16	1
92.	$C_{16}H_{14}N_2O_3$	643568	870	2			1	
93.	$C_{16}H_{14}O_3$	5323699	476	2	4	1	2	2
94.	$C_{16}H_{16}O_2$	5323700	525	1	23	1	2	1
95.	$C_{16}H_{17}NO_2$	U1	1190	3	166	2	7327	6
96.	$C_{16}H_{20}O_2$	U5	410	4	876	1	1	
97.	$C_{16}H_{22}O_{11}$	79064	16	1			1	
98.	$C_{17}H_{14}N_2$	5326090	147	14			5139	1
99.	$C_{17}H_{18}O_3$	5323701	625	1	1		1	
100.	$C_{17}H_{22}O_5$	U17	308	4			5	1
101.	$C_{17}H_{25}NO$	643575	405	6			5	1
102.	$C_{18}H_{16}O_2$	367209	295	2			1	
103.	$C_{18}H_{18}O$	11953648	194	1	1		1	
104.	$C_{18}H_{18}O_3$	U4	487	1	1		1	
105.	$C_{18}H_{18}O_4$	5323702	608	3	1		1	
106.	$C_{18}H_{18}O_4$_b	291963	608	1	816	4	38	3
107.	$C_{18}H_{23}BrO_4$	5326163	10	1			4	1
108.	$C_{19}H_{16}N_4O_4$	U13	380	1	1		1	
109.	$C_{19}H_{16}O$	U2	99	1	443	1	2	1
110.	$C_{19}H_{20}O_2$	5323703	272	5	11	1	1	
111.	$C_{19}H_{32}N_4O_5S_2$	U8	1	1			7	1
112.	$C_{20}H_{17}NO_4$	643578	613	2	2	1	1	
113.	$C_{20}H_{24}O_7$	5324806	141	3			1	
114.	$C_{20}H_{26}NO_3$	5255542	85	2			1	
115.	$C_{20}H_{26}O_3$	94771	334	1	1		1	
116.	$C_{21}H_{22}O_4$	5323713	352	1	1		1	
117.	$C_{22}H_{22}O_6$	U7	336	2	53	1	6	1
118.	$C_{22}H_{26}O_6$	291964	244	1	3	1	11	1
119.	$C_{28}H_{45}NO_8$	U11	11	1	1		1	
120.	$C_{29}H_{36}ClF_5N_2O_4$	609727	1	1			1	
121.	$C_{30}H_{30}O_6$	U16	43	1			5	1
122.	$C_{30}H_{42}ClN_3O_5$	609484	1	1			1	
123.	$C_{35}H_{40}N_2O_4$	U15	26	1	1		1	
124.	$C_{38}H_{55}NO_{10}$	U19	4	1	2	1	1	
125.	Caffeine	2519	141	1			1	
126.	Camphor	2537	660	1			20	4
127.	Chlorfluazuron	91708	3	1			52	1
128.	Clobenzorex HCl	71675	36	2	72	2	1	

Table 3 (*continued*)

#	Dataset	PubChem CID	Structures same MF	^{1}H, MF	^{1}H, $^{13}C_H$	^{1}H, $^{13}C_H$, MF	^{1}H, ^{13}C	^{1}H, ^{13}C, MF
129.	Cortisone	222786	184	26	9	1	1	
130.	Cyclohexanone	7967	191	1			1	
131.	Cyclopentanone	8452	76	1			1	
132.	DDT	3036	6	1			1	
133.	Dihydrotestosterone	15	236	3	6	1		
134.	Dimedone	31358	599	1			1	
135.	Dimestrol	24483	234	1	1		2	1
136.	Ethoxyethene	8023	28	1			1	
137.	Ethyl acetate	8857	61	1			1	
138.	Ethyl cinnamate	7649	596	1			1	
139.	Ethyl sorbate	16970	599	3			1	
140.	Ethylbenzene	7500	87	1			1	
141.	Fexofenadine HCl	3348	36	2	91	1	3	1
142.	Fraxin	5273568	16	2	5	2	2	1
143.	Gibberellic acid	6466	259	6	1		1	
144.	Glycerol acetonide	7528	181	2			1	
145.	Haloperidol	3559	4	1			17	1
146.	Hex-5-en-1-ol	69963	158	1			1	
147.	Hexaphenyldisiloxane	74587	1	1		108	1	
148.	Hydroquinone	785	85	1			1	
149.	Isoindole	305258	516	1			1	
150.	Isopropanol	3776	3	1			1	
151.	Isoquinoline	643561	506	1	1		1	
152.	Isovanillin	12127	158	2			3	2
153.	Juglone	3806	22	1			1	
154.	Lasalocid sodium salt	3887	1	1	7	1		
155.	*m*-Xylene	7929	87	3			6	2
156.	Menthol	1254	305	28	144	1	41	3
157.	Methyl anisate	8499	322	1			1	
158.	Methyl cinnamate	7644	368	1			1	
159.	Nicotine	942	227	4			1	
160.	*o*-Anisidine	7000	167	2			1	
161.	*o*-Xylene	7237	87	2			1	
162.	*p*-Xylene	7809	87	1			1	
163.	Piperazine	3332880	283	31	4	1	1	
164.	Piperine	4840	1332	1			1	
165.	Prednisone	4900	251	17	1		1	
166.	Pulegone	6988	660	1			1	
167.	Pyridine	1049	6	1			1	
168.	Pyrrole	137477	394	1			1	
169.	Quinine	1065	1694	6	1			
170.	Quinoline	7047	12	1			1	
171.	Rescinnamine	32681	12	1			1	
172.	Safrole	5144	368	1			1	
173.	Strychnine	5304	1182	1	1		1	
174.	Strychnine (1 mg)	5304	1182	1	1		1	
175.	Sucrose	1115	44	3	1		1	
176.	Taxol	4666	11	1			1	
177.	Vanillic acid	8468	136	2			2	2
178.	Vanillin	1183	158	2			2	1
179.	Verbenol	61126	660	1	3	1	4	3

3. Results

The NMRanalyst structure identification is evaluated using 179 compounds. Table 3 summarizes the obtained results. The first column in the table numbers the evaluation compounds for easy reference in the subsequent discussion. The second column lists the compound names in alphabetical order. If a compound name is too long, the molecular formula in Hill order is provided instead. Strychnine is included twice in this table (datasets 173 and 174), as it was acquired independently using different sample quantities and different probes. The third table column is the PubChem CID. A "CID" starting with "U" represents a user added structure, which does not exist in the PubChem structure collection. Nineteen structures were added for this study. The fourth column is the total number of FindIt structures with the molecular formula identical to an evaluation compound.

The remaining five columns in Table 3 list the placement (i.e., ranking) of the correct structure under various input combinations of 1D proton (^{1}H), protonated carbon ($^{13}C_H$), and carbon (^{13}C) analysis results, and molecular formula (MF). The major evaluated combinations are proton results plus molecular formula (^{1}H, MF), and proton and carbon analysis results (^{1}H, ^{13}C). The 1D proton spectrum is the fastest routine NMR spectrum to acquire. But its resolution is limited and proton–proton couplings cause an unknown number of contained resonances. It is unlikely to identify the correct structure among millions of candidates based solely on the proton spectral results. Hence the proton information is combined with the molecular formula to obtain practical placements.

Instead of the molecular formula, a molecular weight range can be specified. As the FindIt rankings depend on the accuracy of the specified weight range (with a highly accurate molecular weight likely resulting in a better ranking of the correct structure), the molecular weight input parameter is not evaluated in this study.

Fig. 7 summarizes the percentages with which FindIt identifies the correct structure as the best match (at the place one) for the evaluation compounds. Using the 1D proton spectral results plus the molecular formula (^{1}H, MF) identifies slightly more compounds (64.8%) than using the 1D proton and 1D carbon (^{1}H, ^{13}C) information (63.1%). Combining 1D proton and 1D carbon information with the molecular formula (^{1}H, ^{13}C, MF) achieves the correct structure identification in 89.8% of the cases. The ^{1}H, $^{13}C_H$, MF combination improves the correct identification rate by two percent in this study (91.8%) compared with the ^{1}H, ^{13}C, MF input combination. This is likely due to compromised shift predictions for some quaternary carbons. In general, the ^{13}C spectrum provides more information than the detection of protonated carbons alone.

The 1D proton and protonated carbon (^{1}H, $^{13}C_H$) combination has the lowest correct structure identification rate (49%). The DEPT-135 and HSQC spectra do not observe unprotonated carbons. So the obtained FindIt placements tend to be worse than using the full carbon spectrum information. Poor F1 resolution of an HSQC spectrum can further compromise the FindIt ranking. But for limited sample quantities or when using indirect detection probes, the acquisition of an acceptable 1D carbon spectrum may not be practical. With about a threefold higher sensitivity, the DEPT-135 and HSQC spectra are attractive input options. For practical applications, all available information on a compound should be included for the optimal structure identification.

For the confirmation of the correctness of a proposed structure, a placement within the top 8000 matches from over 8 million candidate structures appears sufficient. This corresponds to a 0.1% probability of falsely classifying a specified structure as the correct one. All evaluation datasets obtain a better FindIt placement under all the evaluated conditions.

4. Discussion

Several evaluation dataset shortcomings can compromise the obtained FindIt placements. The proton spectrum of $C_{10}H_{10}O_2$_c (dataset 58) is specified by WebSpectra with a 9.5 to −0.5 ppm sweep width, so its aldehyde proton around 10 ppm is not observed. The isoindole (dataset 149) and $C_{12}H_{13}N_3O_2$ (dataset 75) shimmings are poor, potentially compromising their methyl/methoxy group detection. The signal-to-noise ratio of the $C_{13}H_{19}N_3O_2$ (dataset 82) and $C_{14}H_{21}N_3O_2$ (dataset 86) carbon spectra is too low for the visual or software detection of all the expected carbon resonances. The carbon spectrum of $C_{16}H_{17}NO_2$ (dataset 95) and of several other samples have spikes around the spectral edges or in the middle of the spectrum. Several datasets contain resonances besides the compound to be determined and the stated locking solvent. $C_{28}H_{45}NO_8$ (dataset 119) contains chloroform in its $CDCl_3$ solvent. $C_{17}H_{14}N_2$ (dataset 98) contains some cyclohexane. Cortisone (dataset 129) and fexofenadine (dataset 141) contain DMSO-d_6 in addition to the specified $CDCl_3$ solvent. Clobenzorex (dataset 128) and fexofenadine (dataset 141) contain maleic acid for obtaining quantitative information. $C_{10}H_{12}O_3$ (dataset 61) and $C_{18}H_{18}O_4$_b (dataset 106) contain a small amount of methanol-d_4. $C_{10}H_{11}NO_4$ (dataset 59) and verbenol (dataset 179) show a strong

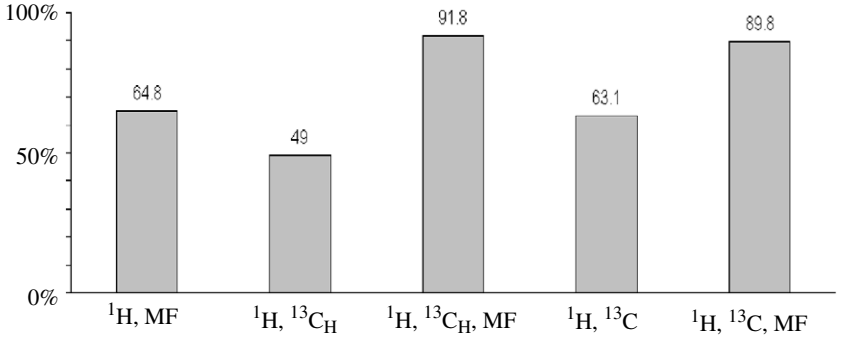

Fig. 7. The percentages with which FindIt identifies the correct evaluation compound structure at place one under various input combinations. The figure summarizes the results from Table 3.

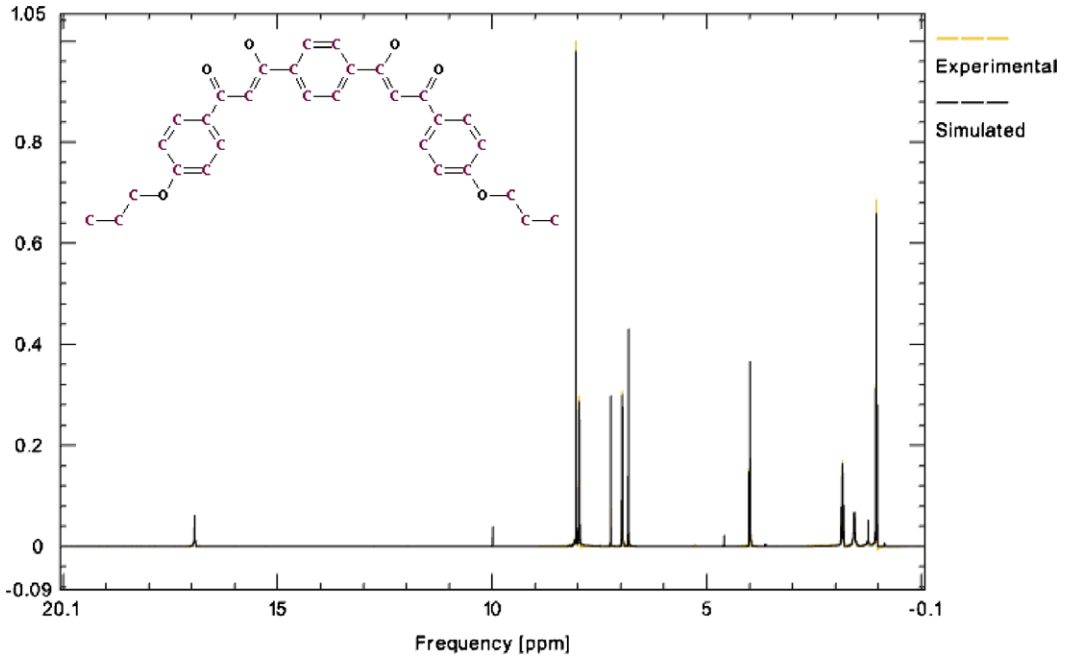

Fig. 8. Structure of evaluation compound $C_{30}H_{30}O_6$ with Bruker AMX 400 MHz acquired 1D proton spectrum. The 16.9 ppm resonance originates from the two enol protons interacting through-space with close-by carbonyl groups.

TMS resonance with detected ^{13}C sidebands. These unexpected resonances were excluded from the analysis results for this evaluation.

The speed of determining best matching structures could be improved using inverted files [19], as demonstrated for carbon shifts by W. Robien's SAHO—Search (http://nmr-predict.orc.univie.ac.at/identify). But the evaluation compound $C_{30}H_{30}O_6$ (dataset 121) proton spectrum, shown in Fig. 8, illustrates a potential limitation. Its 16.9 ppm proton resonance results from both enol protons coordinating with nearby carbonyl groups. For proton data, incorrect referencing, undetected resonances, imperfect shimming, spectral glitches, or shifts greatly affected by through-space interactions, the used exhaustive sequential search is more reliable than a potentially faster inverted file search.

A proton spectrum is expected to contain impurity resonances. But a carbon spectrum analyzed by NMRanalyst is assumed to have neither missing nor additional resonances. Otherwise, a compromised VerifyIt structure consistency rating and subsequent FindIt placement may result. Further work on reducing the dependence on a clean carbon resonance list is in progress.

Currently, only a 20th of HOSE codes with assigned shifts are available for the proton compared to the carbon shift prediction. Heteroatom-bonded proton shifts are not predicted for candidate structures. Stereo chemical information is not utilized for predicting shifts. Gathering more assigned proton shifts from published sources, developing methods for assigning and predicting heteroatom-bond proton shifts, and incorporating stereo chemical information in the shift prediction are planned for the further development of the NMRanalyst structure identification system.

5. Conclusions

A small organic molecule can be identified by matching its proton and/or carbon NMR spectra to the predicted chemical shifts of 8 million candidate structures. The PubChem structure collection includes most published small organic molecules and provides the foundation for this structure identification. When the molecular formula of the unknown is available, a fast to acquire 1D proton spectrum is sufficient to identify the correct molecular structure in 64.8% of the evaluation cases. When the molecular formula is unavailable, a 1D proton and 1D carbon NMR spectra identify the correct structure in 63.1% of the cases. The analyzed spectra and best matching structures for the evaluation compounds are accessible from http://www.sciencesoft.net/FindIt.html.

The structure identification system evaluated in this study is implemented in the NMRanalyst 3.5 software. The software with over 8 million FindIt structures occupies around 1.5 gb of disk space. It can be used on a modern personal computer. Identifying best matching structures usually takes only a few minutes. The software system can be reliably automated and can be used for fast compound identifications, where a full structure elucidation may not be practical. The NMRanalyst software with its data analysis, VerifyIt, and FindIt modules support Varian and Bruker format data and is available for Linux (Red Hat 9, Enterprise 3, & Enterprise 4) and MS Windows (2000 & XP). The web site http://www.sciencesoft.net provides the latest information on the software.

Acknowledgments

This project was made possible by the NIH SBIR Grant R44 MH061652. We greatly appreciate the support and datasets from Dr. Heinz Kolshorn from the University of Mainz, Germany. Additional datasets were contributed by Dr. Dimitris Argyropoulos, Dr. Ronald Crouch, Dr. Péter Sándor, and Dr. Igor Goljer from Varian Inc., Dr. Till Kühn from Bruker BioSpin Switzerland, Prof. Charles G. Fry from the University of Wisconsin—Madison, Dr. Shi Bai from the University of Delaware, Dr. Charles L. Mayne from the University of Utah, the National Institute of Advanced Industrial Science and Technology (AIST), and WebSpectra.

References

[1] T. Lindel, J. Junker, M. Köck, 2D NMR-guided constitutional analysis of organic compounds employing the computer program COCON, Eur. J. Org. Chem. (1999) 573–577.

[2] C. Steinbeck, Recent developments in automated structure elucidation of natural products, Nat. Prod. Rep. 21 (2004) 512–519.

[3] R. Dunkel, C.L. Mayne, R.J. Pugmire, D.M. Grant, Improvements in the computerized analysis of 2D INADEQUATE spectra, Anal. Chem. 64 (1992) 3133–3149;
R. Dunkel, C.L. Mayne, M.P. Foster, C.M. Ireland, D. Li, N.L. Owen, R.J. Pugmire, D.M. Grant, Applications of the improved computerized analysis of 2D INADEQUATE spectra, Anal. Chem. 64 (1992) 3150–3160.

[4] R. Dunkel, Correction and automated analysis of spectral and imaging data, US Patent No. 5,572,125, November 5, 1996;
R. Dunkel, A method for correcting spectral and imaging data and for using such corrected data in magnet shimming, US Patent No. 5,218,299, June 8, 1993; British Patent 0 577 770; German Patent 692 31 690.6-08.

[5] R. Dunkel, C.L. Mayne, J. Curtis, R.J. Pugmire, D.M. Grant, Computerized analysis of 2D INADEQUATE spectra, J. Magn. Reson. 90 (1990) 290–302.

[6] W.F. Reynolds, R.G. Enríquez, Choosing the best pulse sequences, acquisition parameters, postacquisition processing strategies, and probes for natural product structure elucidation by NMR spectroscopy, J. Nat. Prod. 65 (2002) 221–244.

[7] L. Griffiths, R. Horton, Towards the automatic analysis of NMR spectra: Part 6. Confirmation of chemical structure employing both ^{1}H and ^{13}C NMR spectra, Magn. Reson. Chem. 44 (2006) 139–145.

[8] S.S. Golotvin, E. Vodopianov, B.A. Lefebvre, A.J. Williams, T.D. Spitzer, Automated structure verification based on ^{1}H NMR prediction, Magn. Reson. Chem. 44 (2006) 524–538.

[9] H.O. Kalinowski, S. Berger, S. Braun, Carbon-13 NMR Spectroscopy, John Wiley & Sons, 1988, ISBN 0-471-91306-5.

[10] E. Pretsch, P. Bühlmann, C. Affolter, in: Structure Determination of Organic Compounds, third ed., Tables of Spectral Data, Springer, 2000, ISBN 3-540-67815-8.

[11] A. Fürst, E. Pretsch, W. Robien, Comprehensive parameter set for the prediction of the ^{13}C NMR chemical shifts of sp^3-hybridized carbon atoms in organic compounds, Anal. Chim. Acta 233 (1990) 213–222.

[12] E. Pretsch, A. Fürst, W. Robien, 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 248 (1991) 415–428.

[13] R.B. Schaller, C. Arnold, E. Pretsch, New parameters for predicting ^{1}H NMR chemical shifts of protons attached to carbon atoms, Anal. Chim. Acta 312 (1995) 95–105.

[14] U.E. Matter, C. Pascual, E. Pretsch, A. Pross, W. Simon, S. Sternhell, Estimation of the chemical shifts of olefinic protons using additive increments. II. The compilation of additive increments for 43 functional groups, Tetrahedron 25 (1969) 691–697.

[15] J. Beeby, S. Sternhell, T. Hoffmann-Ostenhof, E. Pretsch, W. Simon, Estimation of the chemical shifts of aromatic protons using additive increments, Anal. Chem. 45 (8) (1973) 1571–1573.

[16] W. Bremser, HOSE—a novel substructure code, Anal. Chim. Acta. 103 (1978) 355–365.

[17] W. Bremser, Expectation ranges of ^{13}C NMR chemical shifts, Magn. Reson. Chem. 23 (4) (1985) 271–275.

[18] C. Steinbeck, S. Kuhn, NMRShiftDB—compound identification and structure elucidation support through a free community-built web database, Phytochemistry 65 (2004) 2711–2717.

[19] W. Bremser, H. Wagner, B. Franke, Fast searching for identical ^{13}C NMR spectra via inverted files, Org. Magn. Reson. 15 (2) (1981) 178–187.