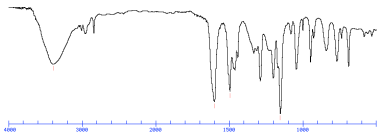
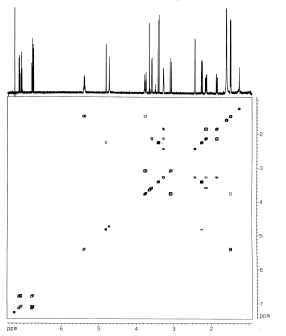
Reviving Analytical Data of the Past with Open Submission Databases and Text Mining Tools

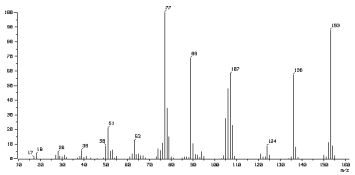
Sam Adams[‡], Stefan Kuhn[†], Peter Murray-Rust[‡*],
Christoph Steinbeck[†*], Joe Townsend[‡]
[‡] Unilever Center for Molecular Informatics, Cambridge, UK
[†] Cologne University Bioinformatics Center (CUBIC), Cologne, Germany

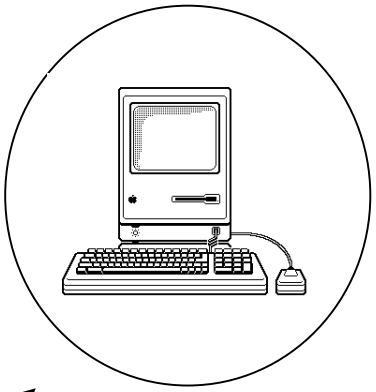
Computer-Assisted Structure Elucidation (CASE)





Molecular Formula from HR-MS



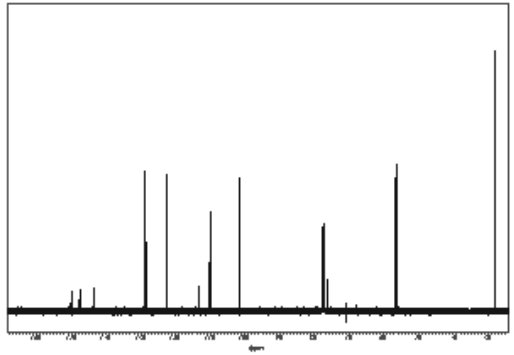


- Constitution

- Conformation
- Chirality
- E/Z Stereochemistry

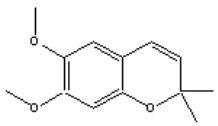
Further Facts:
- Substructures
- Drug Likeness
- Natural Product Likeness
- etc.

Searching spectra (Dereplication)

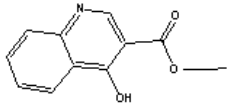


Shift [ppm]
149.99
147.58
143.46
128.64
122.38
113.44
110.01
101.39
76.42
56.90
56.32
28.08

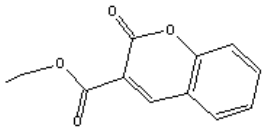
Structure-Spectra Database



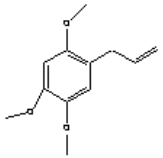
Similarity :99.75%



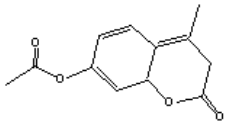
Similarity :58.97%



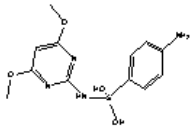
Similarity :53.96%



Similarity :52.67%



Similarity :51.50%



Similarity :51.00%

Steinbeck, C. Computer-Assisted Structure Elucidation. In *Handbook on Chemoinformatics.;* Gasteiger, J. Ed.; Wiley-VCH: Weinheim, 2003; Vol. 2; pp. 1378-1406.

The Situation

- 40 years of research in NMR and even more in other disciplines,
  - Still no open community database for analytical or spectroscopic data.
  - But: Such Databases have become important again in Natural Product Drug Discovery efforts and in Systems Biology
- 40 years full of literature data, waiting to be resurrected.
- Structure and Spectra are partly assigned and partly not.
- Scientists have used their full artistic freedom in layout.
- No semantics for published data, just pixels.

NMRShiftDB

An Open Access, Open Submission, Open Source Database for Organic Molecules and their NMR Data

http://www.nmrshiftdb.org

[1] Steinbeck, C.; *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1733 - 1739.
[2] Steinbeck, C.; Kuhn, S. *Phytochemistry* **2004**, *65*, 2711-2717.

The NMRShiftDB authoring process

Thanks to David Martinsen for the insights regarding the ACS archive – I missed the mid-70s only by chance.

Journal of Organic Chemistry
Table of Contents
Vol. 41, No. 4: February 20, 1976

Feedback | $ Purchase

Stereochemistry, conformations, and carbon-13 nuclear magnetic resonance spectra of 9-phenyl-9-phosphabicyclo[3.3.1]nonane derivatives
John R. Wiseman and Herman O. Krabbenhoft
pp 589 - 593; DOI: 10.1021/jo00866a001

First Page Full: PDF (647K)

Feedback | $ Purchase

Potamogetonin, a new furanoid diterpene. Structural assignment by carbon-13 and proton magnetic resonance
Cecil R. Smith, Richard V. Madrigal, David Weisleder, Kenneth L. Mikolajczak, and Robert J. Highet
pp 593 - 596; DOI: 10.1021/jo00866a002

First Page Full: PDF (555K)

Feedback | $ Purchase

Biomimetic approach to the synthesis of Laurencia metabolites. Synthesis of 10-bromo-.alpha.-chamigrene
Lawrence E. Wolinsky and D. John Faulkner
pp 597 - 600; DOI: 10.1021/jo00866a003

First Page Full: PDF (557K)

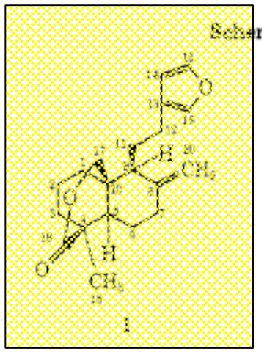
Table I
Selected Proton Chemical Shifts of Potamogetonin, Sciadin, and Nepetaefuran[a]

Protons	Potamogetonin (1)	Sciadin (3)	Nepetaefuran (4)
H-14	6.26, 1 H, dd ($J_{14,15}$ = 1.6, $J_{14,16}$ = 0.9)	6.43, 1 H, dd ($J_{14,15}$ = 1.8, $J_{14,16}$ = 0.9)	6.27, 1 H, dd ($J_{14,15}$ = 1.8, $J_{14,16}$ = 0.9)
H-15	7.33, 1 H, dd ($J_{14,15}$ = $J_{15,16}$ = 1.6)	7.38, 1 H, dd ($J_{14,15}$ = $J_{15,16}$ = 1.8)	7.38, 1 H, dd ($J_{14,15}$ = $J_{15,16}$ = 1.8)
H-16	7.18, 1 H, m	7.47, 1 H, m ($J_{12,16}$ = 0.9)	7.27, 1 H, m
H-17	3.95, 4.15, 2 H, AB q (J = 11)[b]	5.52, d, 1 H (J = 1.3)[c]	4.05, 5.02, 2 H, AB q (J = 12)[d]
H-18	0.87, 3 H, s	1.23, 3 H, s	1.12, 3 H, s
H-20	4.82, 4.94, 2 H, AB q (J = 1)	4.80, 5.05, 2 H, AB q (J = 1)	2.32, 2.70, 2 H, AB q (J = 14)

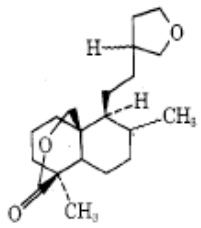
[a] Spectra were determined in $CDCl_3$. Chemical shifts (δ) are expressed in parts per million from tetramethylsilane and couplings (J) in hertz. Letters following the shifts indicate the multiplicities observed without decoupling. [b] The δ 4.15 doublet shows further splitting (J = 1.8 Hz). [c] Irradiation at δ 1.40 eliminates this coupling. [d] The δ 5.02 doublet shows further splitting (J = 1.7 Hz), decoupled by irradiation at δ 2.15.

addition to the expected triglycerides, 25% of an unfamiliar component (1) which subsequently was isolated by thin layer chromatography (TLC).

Scheme I



1



2

By high-resolution mass spectrometry, the empirical formula of 1 was established as $C_{20}H_{28}O_3$. The ^1H NMR spectrum of 1 indicated the presence of a β-substituted furan group (multiplets at δ 6.22, 7.18, and 7.30),[3] an exocyclic double bond (doublets at δ 4.82 and 4.94),[3] and a tertiary methyl group (singlet at δ 0.87). The uv spectrum exhibited a maximum at 201 nm (ϵ 8329), in accord with a furanoid structure.[3,4] The ir spectrum of 1 had maxima at 895 cm^{-1}, suggestive of an exocyclic double bond, and at 872 cm^{-1}, in harmony with a furan nucleus.[3] This ir spectrum also showed a peak at 1740 cm^{-1}, evidently due to an ester or δ-lactone grouping; no hydroxyl absorption was observed.

Catalytic hydrogenation of 1 provided a hexahydro derivative (2), $C_{20}H_{32}O_3$, and thus indicated that 1 probably

data and previously observed biogenetic patterns for furanoid labdane derivatives.[1] The apparent resistance of 1 to alkaline hydrolysis seemingly contradicts our proposed structure, since facile relactonization normally would be expected of a γ-lactone but not of a δ-lactone. Nevertheless, the chemistry of some related diterpenoid δ-lactones —derivatives of columbin,[5] nepetaefolin,[6] and nepetaefuran[6] (4)—provides parallels for the behavior of

trast, sciadin (3) can be saponified and isolated

acid;[7] perhaps its ether linkage makes the corr

free acid relatively stable.

Structure 1 was validated by correlating its ^1H

carbon-13 magnetic resonance (^{13}C NMR) spe

those of two related furanoid diterpenes, sciadi

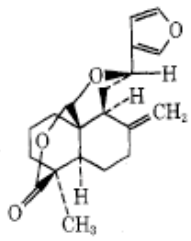
nepetaefuran (4), which were selected as mo

pounds.[6,7] These correlations resulted in assignm

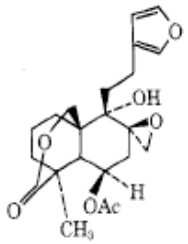
bon-13 chemical shifts for 3 and 4, which had no

amined by ^{13}C NMR previously.

Scheme II



3



4

Table II
Carbon-13 Chemical Shifts of Potamog
Sciadin, and Nepetaefuran[a]

Carbon	Potamogetonin (1)	Sciadin (3)	Nepet
1	41.1 t	40.9 t	41
2	20.9 t	20.8 t	20
3	25.8 t	25.0 t	25
4	33.5 s	37.8 s	32
5	49.5 d	45.0 d	46
6	28.1 t	28.5 t	68
7	36.1 t	35.6 t	42
8	145.2 s	145.1 s	
9	51.7 d	48.4 d	74
10	51.1 s	44.9 s	56
11	23.7 t	33.1 t	32
12	37.0 t	69.4 d	39
13	125.1 s	125.8 s	124
14	110.0 d	108.8 d	110
15	142.6 d	143.2 d	143
16	138.9 d	139.2 d	138
17	76.4 t	100.4 d	73
18	23.7 q	22.3 q	22
19	173.0 s	174.4 s	
20	108.3 t	110.0 t	46
Acetyl			170

[a] Chemical shifts (δ) were measured in parts from tetramethylsilane in $CDCl_3$ solution. [b] N

Experimental Section[18]

General Techniques. ^1H NMR spectra were measured with Varian HA-100 instrument. For ^{13}C NMR spectra, the instrumen was a Varian XL-100 equipped with a Digi-Lab Fourier transfor

ly grouped peaks remains uncertain. Of the th triplets allocated to C-1, C-2, and C-3 in eac pounds, those at highest field are associated cause it has the fewest β substituents and sho farther upfield by the C-17 and C-19 axial They occur 1–2 ppm upfield from the C-2 *trans*-decalin.[9] Differentiation of the peaks fo is tentative and is based on assignments b Stothers[9] for *trans*-decalin systems.

Two singlets at high field, from C-4 and C- pated in the spectrum of 4. That of C-10 s lower of the two because of the larger number ents. The same order can be retained for 1 an the following correlations: the 6β-acetoxy gro

Chemical Archeology

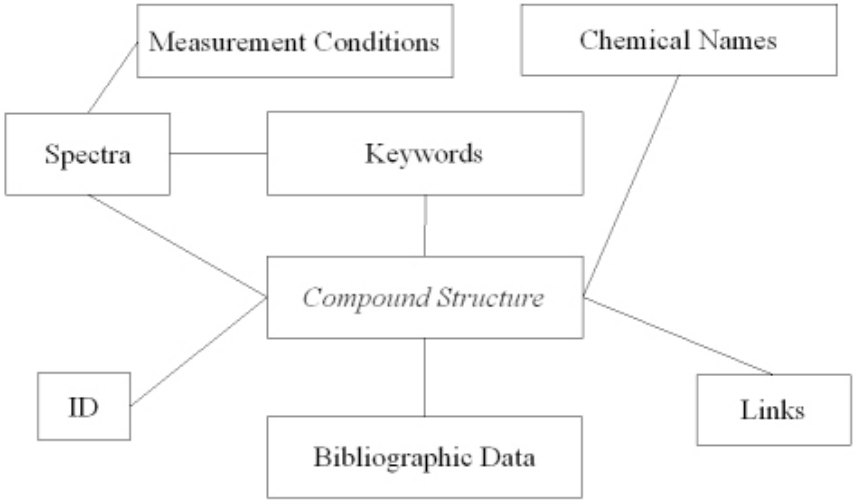
Entering all your discovered information into a web form

Show Movie. Don't forget it. Show it now!

NMRShiftDB (http://www.nmrshiftdb.org)

An open access, open submission, open source database for organic structures and their NMR data

- >19.000 Structures and > 21.000 assigned spectra
- Searching for spectra, structures and facts
- NMR Prediction for ^{13}C-, ^{1}H-NMR and other nuclei
- Corina-generated 3D coordinates
- ACD/Names^{TM} IUPAC-Names
- Open Submission, Peer-Review by human reviewers

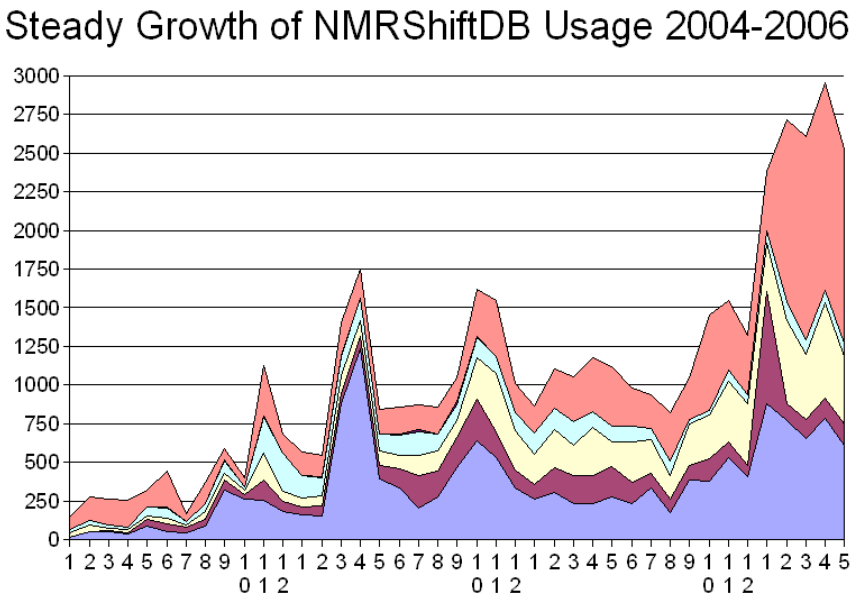


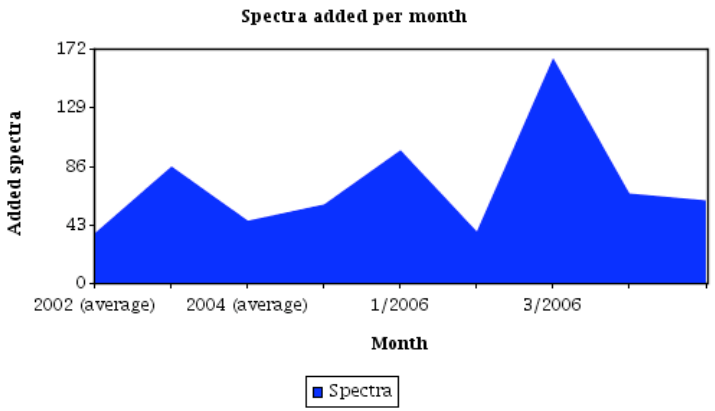
Steinbeck, C.; Kuhn, S.; Krause, S. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1733 - 1739.
Steinbeck, C.; Kuhn, S. *Phytochemistry* **2004**, *65*, 2711-2717.

NMRShiftDB (http://www.nmrshiftdb.org)

An open access, open submission, open source database for organic structures and their NMR data

- 900 registered users
- Runs on 4 mirrors servers in three geographic locations, two more soon





Steinbeck, C.; Kuhn, S.; Krause, S. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1733 - 1739.
Steinbeck, C.; Kuhn, S. *Phytochemistry* **2004**, *65*, 2711-2717.

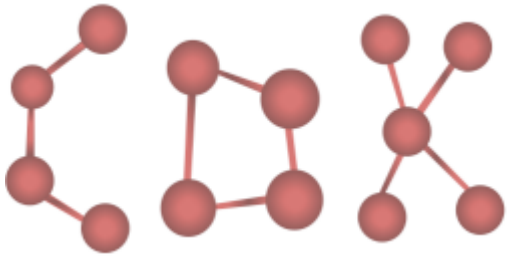
Towards Automated Chemical Archeology

Yet another demo...

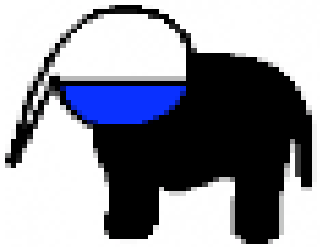
Outlook

- Database section with structure-spectrum relations without assignment – open to later assignment by community.
- Work on articles scanned with OCR.
- Recognition of connection tables from chemical structure bitmaps by Optical Chemical Entity Recognition (OCER).
- Open Editor-Software for semiautomatic semantic annotation of old articles.

Acknowledgment



- Bioclipse team
- CDK team
- CML team
- Apache team, Linux Torvalds and the Open Source Community as a whole
- .... and in alphabetical order:
  - Peter Corbett
  - Tobias Helmus
  - Stefan Kuhn
  - Peter Murray-Rust
  - Ola Spjuth
  - Egon Willighagen



OSCAR 3

- ... as well as the German Research Council (DFG) for generous funding

URLs

http://cdk.sourceforge.net (Chemistry Development Kit, CDK)
http://www.chemistry-development-kit.org (Showcase-Site)
http://www.nmrshiftdb.org (NMR-Database)
http://www.bioclipse.net (Bioclipse Development Site)
http://www.blueoblisk.org (Blue Obelisk Movement)