To add or not to add—that is the question for your reference spectra prediction databases

Tony Davies,[a] Wolfgang Robien[b] and Jeff Seymour[c]
[a]External Professor, University of Glamorgan, UK and Informatics Consultant
[b]Institute for Organic Chemistry, University of Vienna, Austria
[c]ModGraph Consultants, 1 Oakland View, Welwyn AL6 0RJ, UK

Having a good idea

It is generally accepted that spectra prediction performance can be greatly enhanced if you add your own reference quality data to the database used for the predictions. This is mostly due to the fact that the reference databases that are generally available often represent only a small selection of data from across chemistry. What is probably worse, some prediction systems consist of data files that have been donated by one or two major industrial chemical manufacturers and as such are slanted towards the particular chemistry practiced within those companies—such as large organics or heterocyclic chemistry. They can therefore predict with excellent results within those fields but fall down badly when asked to predict spectra for structures whose chemistry is completely different.

So just buying bigger and bigger databases may not be the way forward to improving your prediction quality. Overall system performance is not really an issue anymore. Running Hierarchical Ordered description of the Substructure Environment (HOSE) code prediction data for medium sized structures C10, C20 against 100,000 in your reference database takes about 1–2s, using neural networks reduces this to a few ms. What is critical is to add reference data representative of your own chemistry. But, with limited resources available, which data should be added to really make a difference?

Recently, Wolfgang Robien and his colleagues were discussing this problem during a congress, with the example of a large pharmaceutical company who were looking to exploit their large collection of ^{13}C-NMR spectra by making it available within their organisation for spectra prediction.

They came up with the idea of using the CSEARCH prediction engine[1] itself to locate the reference data sets of most value to improving the prediction quality of the knowledge base.

Avoiding unnecessary laborious spectra/structure assignments

Now if we look at the 100,000 spectra which represent the candidate data for adding to the knowledge base, we will usually find that most of the data sets have not been assigned. This means that there is no electronic assignment of the spectra features to the registered chemical structures. Were they all available as assigned data it would be possible to simply add them to the prediction set—and monitor statistically the prediction performance of the system.

But this is not how the real world works, so it is important to decide how to deploy limited resources to the best effect. The key question which needs to be answered is how do I select which of the candidate spectra and structures should be added to the database in order to maximise the benefit? In other words how much assignment work do I need to carry out to achieve the best return on my investment?

HOSE codes

Now to understand how their solution works you need to understand something about how NMR prediction software does its job. Back in 1978 Wolfgang Bremser, then of BASF, published a paper "HOSE—A Novel Substructure Code" describing an encoding system for chemical structures which was ideally suited for use as a descriptor in NMR spectra prediction—the HOSE code.[2] HOSE codes describe the structural neighbours of the particular atom of interest, which in NMR essentially identifies those atoms within the molecule influencing the chemical shift of that atom. The neighbouring atoms are described in spheres—the nearest neighbours being in the first sphere. Typically, prediction databases only went out as far as five (or six) (CSEARCH, NMRPREDICT and NMRBENEFIT use up to five) spheres. Figure 1 shows how much of the testosterone molecule would be encoded for a central atom of interest.

In the reference database to be used as the knowledge-base for the predictions, the ^{13}C NMR reference chemical shifts are stored along with HOSE codes

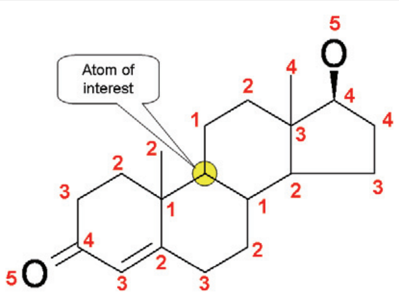


Figure 1. HOSE codes describe nearest-neighbour relationships to atoms of interest. In this figure the atoms are labelled with the sphere number they have in relation to a particular atom of interest. The larger the number of a particular atom, the further away that atom lies from the atom whose NMR chemical is being predicted and probably the weaker the influence of that neighbour to the chemical shift.

for all spheres from one to five. A database with 100,000 reference spectra and their accompanying structures will yield (usually) more than 1 million structure codes. This is the knowledge base which the ^{13}C-NMR prediction algorithms draw upon to calculate new chemical shift values for our chemical structures.

Using HOSE codes for prediction

How does the prediction work? For our atom of interest, the HOSE code for the first sphere is generated and searched for in the reference database. If it is found then the HOSE code going out to the second sphere is searched for. Normally this continues, if possible, until the fifth sphere has been reached.

The prediction application then averages the chemical shift values it has mined from the reference database for the largest number of spheres and presents this as predicted value. Obviously various software packages have numerous refinements on how the data is treated and the predictions weighted, but the basic principles are essentially the same.

For ^{13}C-NMR, the predictions can be extremely reliable as the HOSE structure encoding methodology is exceptionally close to the real physical effects acting to change an atoms NMR shift value. Figure 2 shows an example of one such prediction package displaying the number of spheres used for the various chemical shift predictions so that the user can make their own evaluation about the quality of the predicted values.

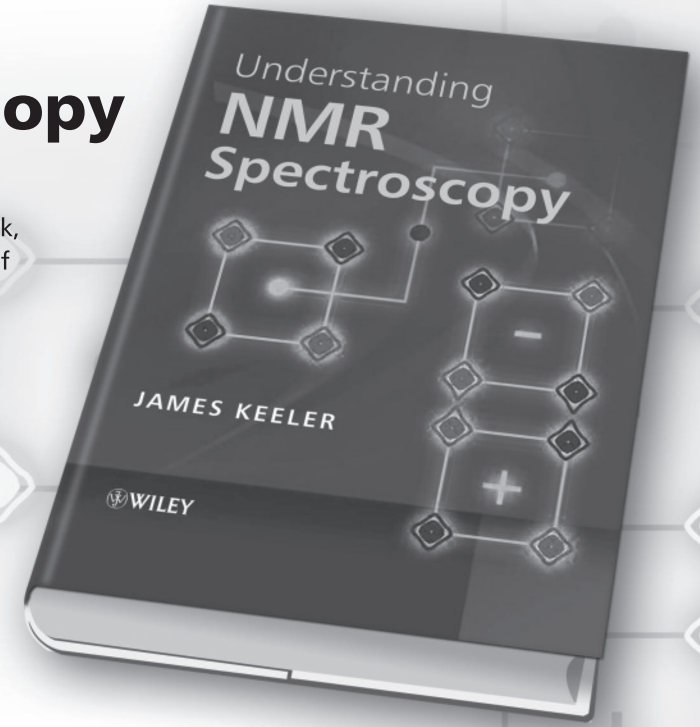
How best to improve predictions

As you can see, if your reference database contains different chemistries you will have little or no chance of locating

The Science of Spectroscopy

Understanding NMR Spectroscopy

JAMES KEELER, University of Cambridge



This authoritative and comprehensive book, allows the reader to build on the theory of NMR and deepens their understanding of how modern high-resolution NMR experiments actually work.

Based upon popular lectures by the author, the text works through familiar ideas and elaborates as the discussion progresses. Sufficient quantum mechanics is introduced to enable a proper analysis of the pulse sequences, but the approach taken is relaxed. All calculations are worked through step-by-step, and a commentary is given at each stage so that the key concepts and ideas introduced are easily digested. Features end of chapter exercises with an online solutions manual.

SpectroscopyNOW website: http://www.spectroscopynow.com/nmr

0470017864 • October 2005 • Hardback • 476 pages
£85.00 • €129.00 • $155.00

FASTLINK / CIRCLE 014 FOR FURTHER INFORMATION

HOW TO ORDER...

EUROPE, MIDDLE EAST,
& AFRICA
John Wiley & Sons Ltd
Tel: +44 (0)1243 843294
Fax: +44 (0)1243 843296
E-mail: cs-books@wiley.co.uk
www.wiley.com

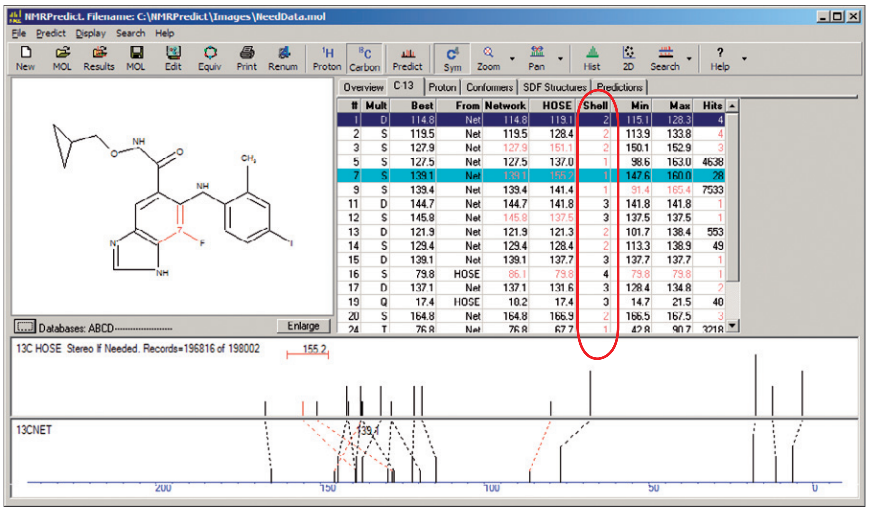
NORTH, CENTRAL &
SOUTH AMERICA
John Wiley & Sons Inc
Tel: 877 762 2974
Fax: 800 597 3299
E-mail: custserv@wiley.com
www.wiley.com

GERMANY, SWITZERLAND
& AUSTRIA
Wiley-VCH Verlag GmbH
Tel: +49 6201 606 400
Fax: +49 6201 606 184
E-mail: service@wiley-vch.de
www.wiley-vch.de

All books are available from your bookseller.
Prices subject to change.
Postage and handling additional.

WILEY

7992 (10/05)


File Predict Display Search Help
New MOL Results MOL Edit Equiv Print Renum
¹H Proton ¹³C Carbon Predict Sym Zoom Pan Hist 2D Search Help
Overview C-13 Proton Conformers SDF Structures Predictions
NH
CH₃
NH
N
F
NH
I
Enlarge
155.2
139.1
200
150
100
50
0

NMRPredict. Filename: C:\NMRPredict\Images\NeedData.mol

#	Mult	Best	From	Network	HOSE	Shell	Min	Max	Hits
1	D	114.8	Net	114.8	119.1	2	115.1	128.3	4
2	S	119.5	Net	119.5	128.4	2	113.9	133.8	4
3	S	127.9	Net	127.9	151.1	2	150.1	152.9	3
5	S	127.5	Net	127.5	137.0	1	98.6	163.0	4638
7	S	139.1	Net	139.1	155.2	1	147.6	160.0	28
9	S	139.4	Net	139.4	141.4	1	91.4	165.4	7533
11	D	144.7	Net	144.7	141.8	3	141.8	141.8	1
12	S	145.8	Net	145.8	137.5	3	137.5	137.5	1
13	D	121.9	Net	121.9	121.3	2	101.7	138.4	553
14	S	129.4	Net	129.4	128.4	2	113.3	138.9	49
15	D	139.1	Net	139.1	137.7	3	137.7	137.7	1
16	S	79.8	HOSE	86.1	79.8	4	79.8	79.8	1
17	D	137.1	Net	137.1	131.6	3	128.4	134.8	2
19	Q	17.4	HOSE	10.2	17.4	3	14.7	21.5	40
20	S	164.8	Net	164.8	166.9	2	166.5	167.5	3
24	T	76.8	Net	76.8	67.7	1	42.8	90.7	3218

Databases: ABCD

13C HOSE Stereo If Needed. Records=196816 of 198002

13CNET

Figure 2. A screenshot from a ^{13}C-NMR prediction software package called NMRPredict showing (circled in red) how many spheres, or shells in this case, found in the reference database have contributed to the predictions.

reference shift values with the three, four or five spheres identical to the atom to be predicted. So now you know why it is better to have more, similar, reference data.

Robien's idea was beautiful in its simplicity and essentially turned the prediction route on its head. To identify which of the new data sets it would be worthwhile adding to the reference database he took the chemical structure of each of the new reference data sets and searched the database as if he was looking to predict its spectrum. By looking at the number of spheres found in the database it was possible to identify those structures which would have been well predicted by the current reference data and those structures which would have been poorly predicted, i.e. those structures whose encoded HOSE codes were not well represented in the database. By assessing which molecules would add the most new HOSE codes to the reference database you get a clear indication which reference data sets it would be worthwhile carrying out the time consuming assignment work on. More mathematically:

- you start with N candidates for selection
- you do a prediction for all of them against your existing database holding X entries
- you select the structure which is represented worst by your reference collection
- you "virtually" add this structure to your reference collection (has now $X+1$ references) in order to benefit already from this dataset for the next prediction
- now you have $(N-1)$ candidates and go back to step #1 until all candidates have been processed.

You end up with a detailed list in which sequence your internal data have to be assigned in order to get most benefit for future predictions.

So the end result is you mine your current reference database against your available additional reference data to come up with a sound basis for deciding how much effort will produce the best return on investment on improving NMR structure predictions.

For more information on NMRBenefit see Reference 3.

References

1. H. Kalchhauser and W. Robien, *J. Chem. Inform. Comput. Sci.* **25**, 103–108 (1985).
2. W. Bremser, *Anal. Chim. Acta* **103**, 355–365 (1978).
3. http://www.modgraph.co.uk/product_nmr_benefit.htm