

A head in the clouds?—Part two: exploring distributed, multi-server ^1H NMR prediction

Antony N. Davies,^{a,b} Mohan Cashyap,^c Robert Lancashire^d and Robert M. Hanson^e

^aStrategic Research Group—Measurement and Analytical Science, Akzo Nobel, Deventer, the Netherlands

^bSERC, Sustainable Environment Research Centre, Faculty of Computing, Engineering and Science, University of Glamorgan, UK

^cMASS Informatics, Harpenden, UK and Tech Mahindra, London, UK

^dDepartment of Chemistry, University of the West Indies, Mona Campus, Kingston, Jamaica

^eDepartment of Chemistry, St Olaf College, Northfield, Minnesota, USA

Distributed ^1H NMR prediction—clouds across the Atlantic...

In the last column we looked at the basics of what is expected of a “cloud”-based system and appealed for information on suppliers of systems of use to spectroscopists that could loosely fit into this category. In August the IUPAC meeting of the Committee on Publications and Cheminformatics Data Standards (CPCDS) took place at de Gruyter’s offices in Berlin over the weekend of 26/27 July. This was opportune as Professor Robert Lancashire attended in his capacity as Titular Member. So after a bit of a long drive we were able to meet up for some good food and a chat about work that he has recently been involved in. Robert was one of the first to respond to the appeal for information following the article. What Robert was so enthusiastic about and presented at the meeting was developed in an intensive collaboration with Bob Hanson of the Department of Chemistry at St Olaf College in Northfield, Minnesota in the USA (<http://www.stolaf.edu/people/hansonr/>).

Together they had developed a web page linking services from multiple sites on both sides of the Atlantic to deliver molecular structure drawing, 3D structure representation, name to structure conversion and display of data generated by linking chemical structures to

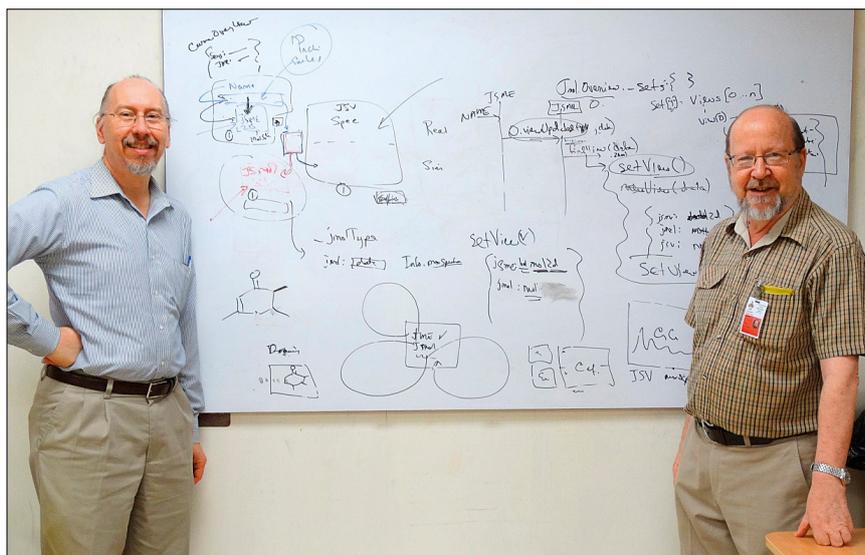


Figure 1. The two Roberts, Hanson to the left and Lancashire to the right. Chemometricians amongst you will clearly spot the false correlation between the name Robert and a particular type of beard!

Luc Patiny’s ^1H NMR prediction engine (Ecole Polytechnique Fédérale de Lausanne, <http://cheminformatics.epfl.ch/>) all available through a single web page. Now this system clearly met the category of software as a service (SaaS) cloud service model and as a public cloud deployment model as it clearly meets the criteria of “The cloud infrastructure is provisioned on the premises of the cloud provider for open use by the general public. It may be owned, managed, and operated by a business, academic, or government

organization, or some combination of them.” as discussed in the last issue. The essential features of on-demand self-service, broad network access and resource pooling are also met, but the only features missing from the NIST definition¹ were essentially commercial in nature around rapid elasticity and measured service although if this were a commercial development I have no doubt that these would be simple to deliver.

Anyway, enough theorising, let us look at the system in operation.

TONY DAVIES COLUMN

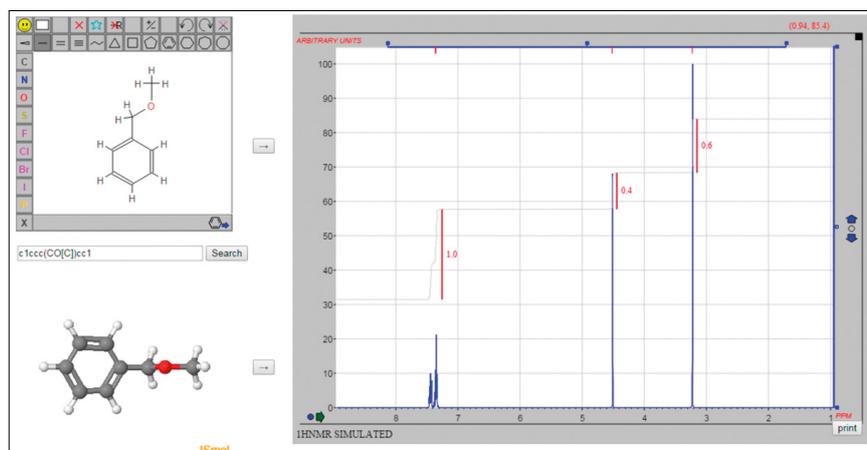


Figure 2. Starting with a drawn structure.

Getting the structure right... What's in a name?

Mohan was good enough to play with what we should probably call a "service". This was clearly an interesting experience for him coming from the pharma industry, where the high level of system scrutiny by regulators usually makes anything which even slightly approaches an "open" system in the eyes of the FDA something to be avoided at all costs due to the difficulties in proving compliance to 21 CFR part 11 and the predicate rules! You can access the site (either through Robert Lancashire's site at http://wwwchem.uwimona.edu.jm/spectra/jsmol/HNMR_predict.html in Jamaica or with possibly better access speeds depending on the time of day at Bob Hanson's site at http://chemapps.stolaf.edu/jmol/jsmol/jsv_predict2.htm).

In our simple example (Figure 2) we have actually started with a drawn structure using the enhanced JSME molecular structure editor by Peter Ertl and Bruno Bienfaitame, Novartis Institutes for BioMedical Research.² The newer version of the Jmol chemical structure editor is upgraded to allow it to also run on iOS-based systems such as iPad or mobile phones. It might be worth noting that if you draw a non-existent molecule this arrangement will still predict a spectrum. This distinguishes it from database lookups. Of course the largest reference database of HNMR spectra is still only tiny compared to the number of known molecules.

Mohan pushed the system further, starting using the search term "Amoxicillin" typed into the text box to test out the link to NCI's Chemical Identifier Resolver server in Frederick, Maryland, USA (<http://cactus.nci.nih.gov/chemical/structure>). The link worked fine and returned a 2D chemical structure (the call looks like: <http://cactus.nci.nih.gov/chemical/structure/>"structure identifier"/"representation"). This 2D file is then used as the input stream in a call across the Atlantic to the input side of the nmrdp system at Lausanne, Switzerland, which carries out the ¹H NMR predic-

tion using SPINUS (Structure-based Predictions In Nuclear magnetic resonance Spectroscopy), a neural network based approach.³⁻⁵ A 3D structure is generated by the CORINA algorithm (Table 1).

Bob Hanson put in an enormous amount of work to make the interfaces talk the correct language to each other and this included writing specific atom numbering code. A major challenge in this integration involved the fact that 2D SDF files do not have the same atom numbering as 3D SDF files, and, to make it worse, the 2D SDF files sent to Lausanne were not the same as those used in and returned from the SPINUS calculation. The solution was to use the SMILES capabilities of Jmol to correlate atoms in the 2D representation with atoms in the 3D representation as well as the returned peak assignments.⁶

With the atom numbering now remaining consistent between each of these steps, the process which allows the final step—the display of the predicted spectrum back in the window in the JSpecView⁷ applet—was complete. The applet from Robert Lancashire's team displays the predicted NMR data in such a way that all of the peak integration and peak identification between the other applet windows still function

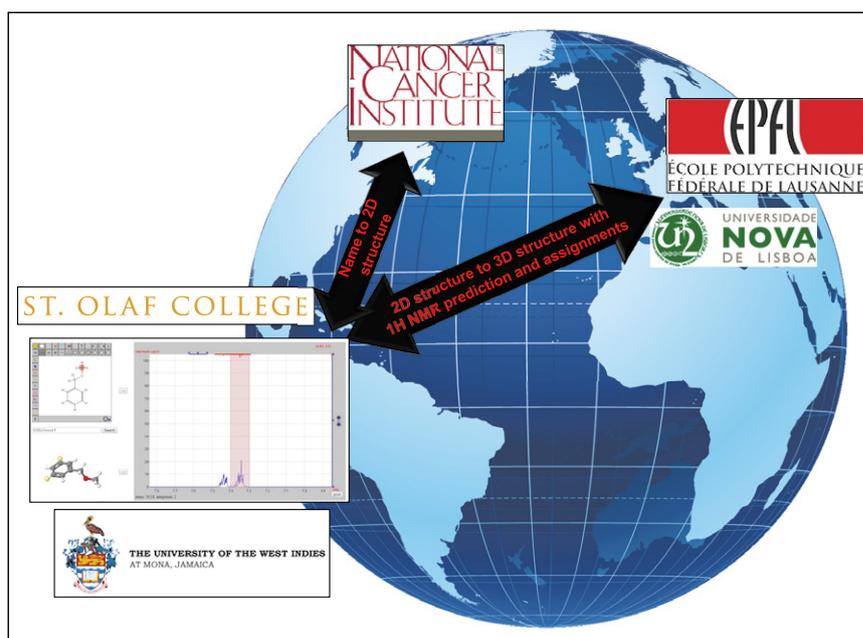


Figure 3. The journey in the cloud across the Atlantic.

TONY DAVIES COLUMN

Table 1. Path of information exchange starting from a structure and from a name (see also Figure 3).

from drawing the structure:

- 1) JSME \Rightarrow SMILES (no H atoms)
- 2) SMILES \Rightarrow NCI \Rightarrow 2D SDF file (with H atoms)
- 3) 2D SDF file \Rightarrow JSME for display
- 4) SMILES \Rightarrow NCI \Rightarrow 3D SDF file
- 5) 3D SDF file to Jmol
- 6) 3D mol file to EPFL gets spectral data, assignment and (modified) 3D mol file

from a name:

- 1) name \Rightarrow NCI \Rightarrow 2D SDF file (with H atoms)
- 2) 2D SDF to JSME
- 3) name \Rightarrow NCI \Rightarrow 3D SDF file
- 4) 3D SDF file to Jmol
- 5) 3D mol file to EPFL gets spectral data, assignment and (modified) 3D mol file

EPFL: École Polytechnique Fédérale de Lausanne

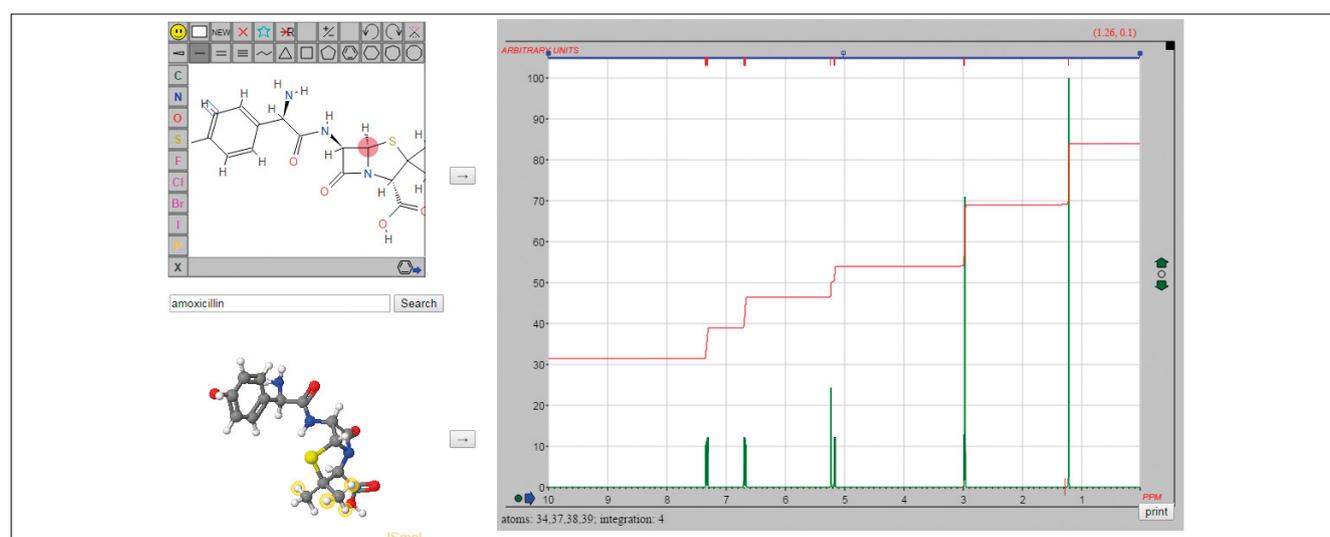


Figure 4. Clicking on a peak in the NMR rendition highlights the responsible atoms in the structure display windows.

as originally designed. This means that if you click on a particular peak in the predicted spectrum from Lausanne the atoms in the 2D and 3D molecular structure renditions are highlighted, for example.

Potential weakness of complex cloud solutions

Now, all the advantages of the speed of the modern internet connections allow the data generation and transfer between multiple servers on various sides of the Atlantic to be completed in just a few seconds... when the servers are actually running. Unfortunately, we do not have any statistics on the availability of the various servers as a team, but clearly when any one of the systems in the chain is down for maintenance or overloaded, the system fails; which can be rather embarrassing if you are in the middle of a presentation.

One of my worries when we decided to feature this initiative in this column was whether when we came to write the column the systems would actually all be running. To my delight they have shown no sign of any stress during our testing—the only somewhat annoying and confusing error message which we got was when I tried to repeat Mohan's amoxicillin experiments and the NCI server responded telling me it was unavailable—when it actually meant “you cannot spell amoxicillin”! Clearly if this were to be commercialised the availability of the servers would be hardened and comply with standard commercial service level agreements. All in all it has been great fun playing with this system as a somewhat complicated example of distributed “cloud” services and I hope Bob and Robert can be persuaded to continue the good work... and other not-for-profit

organisations open up their systems by providing web service functionality such as used here.

References

1. P. Mell and T. Grance, *The NIST Definition of Cloud Computing*. NIST Special Publication 800-145 (2011). <http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>
2. B. Bienfait and P. Ertl, “JSME: a free molecule editor in JavaScript”, *J. Cheminform.* **5**, 24 (2013). doi: <http://dx.doi.org/10.1186/1758-2946-5-24>
3. D. Banfi and L. Patiny, “www.nmrdb.org: Resurrecting and processing NMR spectra on-line”, *Chimia* **62**, 280–281 (2008). doi: <http://dx.doi.org/10.2533/chimia.2008.280>
4. A.M. Castillo, L. Patiny and J. Wist, “Fast and accurate algorithm for the simulation of NMR spectra of large spin systems”, *J. Magn. Reson.* **209**, 123–130 (2011). doi: <http://dx.doi.org/10.1016/j.jmr.2010.12.008>
5. J. Aires-de-Sousa, M.C. Hemmer and J. Gasteiger, “Prediction of ^1H NMR chemical shifts using neural networks”, *Anal. Chem.* **74**, 80–90 (2002). doi: <http://dx.doi.org/10.1021/ac010737m>
6. <http://sourceforge.net/projects/jsmol/>
7. <http://sourceforge.net/projects/jspecview/>