OBC
www.rsc.org/obc

Experimental data checker: better information for organic chemists

S. E. Adams,[a] J. M. Goodman,*[a] R. J. Kidd,*[b] A. D. McNaught,[b] P. Murray-Rust,*[a] F. R. Norton,[a] J. A. Townsend[a] and C. A. Waudby[a]

[a] *Unilever Centre for Molecular Informatics, Department of Chemistry, Lensfield Road, Cambridge, UK CB2 1EW. E-mail: jmg11@cam.ac.uk; E-mail: pm286@cam.ac.uk; Fax: +44 1223 336362; Tel: +44 1223 336434*

[b] *The Royal Society of Chemistry, Thomas Graham House, Science Park, Milton Road, Cambridge, UK CB4 0WF. E-mail: kiddr@rsc.org*

Received 29th July 2004, Accepted 2nd September 2004
First published as an Advance Article on the web 30th September 2004

An experimental data checker has been developed that reads, analyses, and cross-correlates experimental information copied and pasted from authors' manuscripts, which will be useful for authors, referees, editors and readers of papers reporting new molecular information, and which makes possible a quantification of the accuracy of journals' data.

Experimental data on new molecules in organic and inorganic chemistry is presented in a standard form, which varies little from journal to journal. Typically, the appearance of the compound is described, followed by melting point (if applicable), R_f, infra-red and NMR data, and mass spectral information. Other information may also be available. This information is manually summarised from the original experimental data, and there is an opportunity for errors and inconsistencies to be introduced.

We have developed a Java applet, which helps address this problem, and will help improve the quality of published experimental data. Either whole papers, or individual paragraphs of experimental data, can be copied from a word-processing application and pasted into the applet, and analysed interactively. The applet analyses the text, reports problems with parsing the data, and issues from cross-correlating the data. The applet uses Java 1.1 and has been tested on a range of PCs, Apple Mac and Linux computers. The applet is straightforward to use, and, although it does not analyse all of the data that is available in most papers (currently, for example, it discards all diagrams), it analyses enough to provide useful feedback. Its aim is not to provide an exhaustive analysis of all aspects of the data nor to certify a paper correct in all respects. It can provide some helpful information, which an author can use to improve the paper, a referee can use to check the paper, and a reader can use to analyse the paper.

For example, the experimental data† for amphidinoketide I, (Fig. 1), which has recently been reported,[1] was copied from the paper, and pasted into the applet. The result is illustrated in Fig. 2 for the *Safari* browser running in Apple MacOSX. The program has also been tested on a variety of browsers on Windows and Linux systems. This process provides an extremely simple user interface, but the simplicity has a downside. Formatting information is not reliably transfered and all figures are lost. Despite this, the applet is able to parse and analyse the data describing the molecule. Once the data is pasted into the applet, the *Go* button is clicked. The textual data is then analysed, usually within a few seconds if data for only a small number of molecules have been provided, and a variety of displays are available. Clicking on *Details* gives an analysis of the data which has been found. Fig. 3 shows a part of this.

The applet has decided, correctly, that the data that was pasted describes just one molecule in this example. If there were several molecules, the buttons at the bottom of the window could be used to move between them. The first part of the ^{13}C NMR data is shown in the main window, listed as a table. Scrolling down in the window gives all the other data which has been identified for this compound. Even though all figures are lost in the pasting process, the applet picked out the ^{13}C and ^{1}H NMR data, the mass spectrometry, the infra-red data, the R_f and the optical rotation.

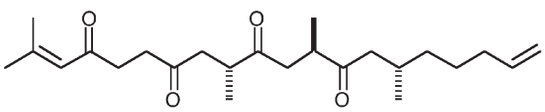

Fig. 1 Amphidinoketide I.

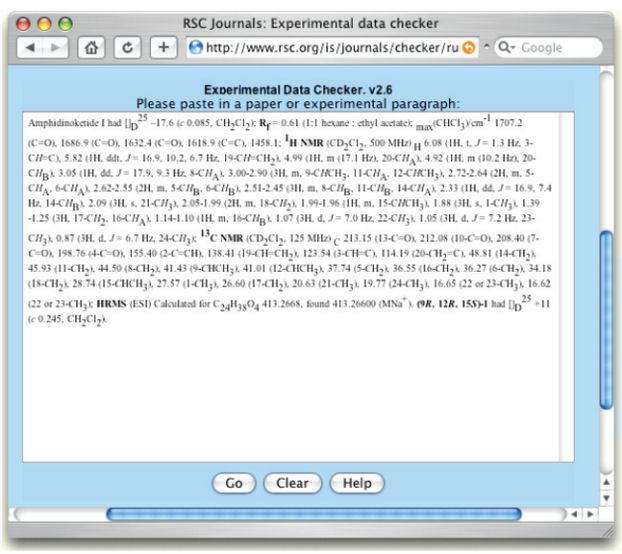

Fig. 2 The applet with amphidinoketide data.

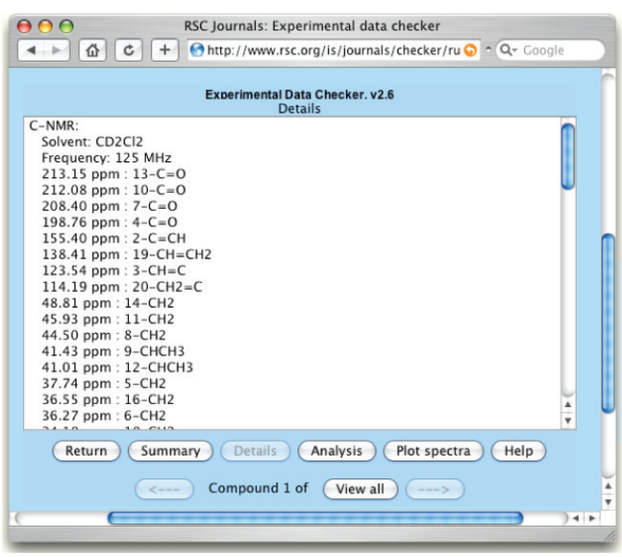

Fig. 3 Details of the analysis.

DOI: 10.1039/b411699m

Clicking on the *Analysis* button now displays comments on the molecular data (Fig. 4). The most important issue is that the HRMS peak does not match the reported value, but differs by about 23, the relative atomic mass of sodium. Checking the experimental data for the HRMS shows that the data is reported for MNa^+ and not M^+. Clicking *Return*, and adding an extra Na to the molecular formula, removes this error. We hope that this process will be automatic in a future version of the program. Three ^{13}C NMR peaks are highlighted as being rather high—over 200 ppm. Such values are slightly unusual, and so are worth checking, but in this case they correspond to ketones, which may have such values. The results were cross-correlated and no other issues were found, although minor issues are noted for the proton and infra-red data.

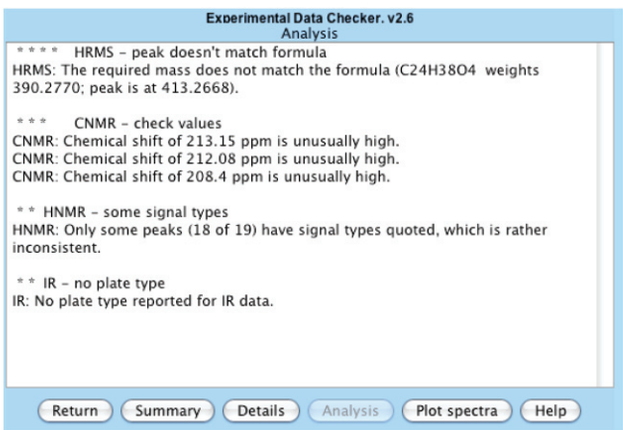

Fig. 4 Comments on the amphidinoketide I data.

The applet calculated that the molecule has six double bond equivalents and checked that the reported HRMS value correlated both with the calculated value given in the data, and with a value re-calculated from the molecular formula. Had any of these tests failed, the applet would have generated error messages, or warnings and comments for less serious issues. The applet also lists any part of the data that has not been analysed. This can make it clear if the data is missing a semi-colon, or has unpaired brackets, or some other typographical problem.

The ^{13}C NMR spectrum for the molecule was reconstructed by the applet (Fig. 5). The buttons list the other spectra which may be displayed, and also allow peak labels to be turned on and off. Fig. 5 shows just one label, the peak that the mouse had last been over when the screenshot was taken.

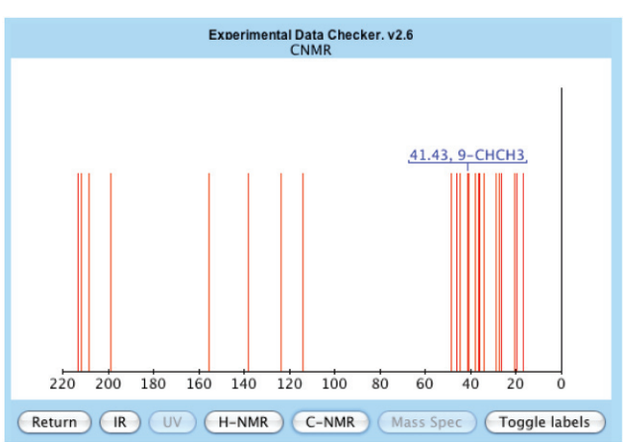

Fig. 5 ^{13}C NMR spectrum reconstructed from experimental data.

As well as checking published data, the applet can also be a useful aid in writing papers, reports and theses. Draft experimental details can be pasted into the applet, as they are being written, and analysed almost instantaneously, providing a straightforward check for correct formatting and for typographical errors. All of the data pasted into the applet is processed on the local computer, and nothing is passed back to the originating site.

The source code, under an open source Artistic License,[2] is also available. The Java program is organised as a toolkit in a series of independent classes, and so can be easily reused in new programs. The same toolkit has been used to create a Java application, which can also be downloaded from the website and run on any computer which has the Java runtime environment installed. This has been illustrated in Fig. 6, with the same data as before. Although the application parses the data in exactly the same was as the applet, and so should produce the same analysis, it is able to do some things that the applet cannot, including highlighting the data in different colours, as shown in Fig. 6. It is now clear at a glance how much of the data has been analysed, and into what groups. ^{13}C NMR, and ^{1}H NMR data, *etc.*, are all coloured differently, and all the data is coloured, except for the final phrase, which refers to a different molecule and so is correctly ignored, and the name of the molecule, which is not recognised as a chemical name in this case, as it is not a IUPAC-style name.

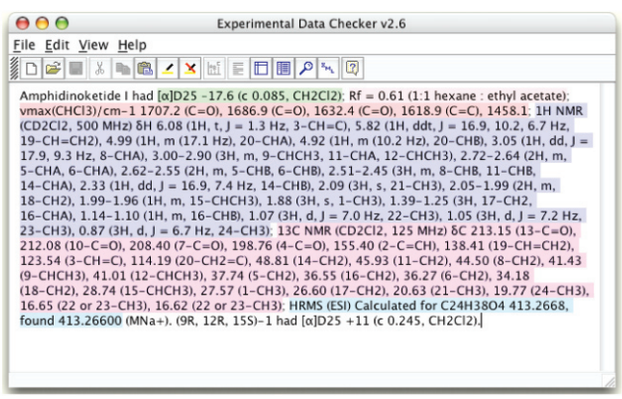

Fig. 6 Java application with the amphidinoketide I data.

If errors are introduced into the data, the display changes. For example, Fig. 7 shows the same results as Fig. 6, except that the proton NMR data is now described using the solvent CD_2Cl_3 measured on a 500 THz spectrometer, a comma has been changed to a semicolon in the middle of the proton NMR data, and the carbon NMR data is now given for ^{12}C rather than ^{13}C.

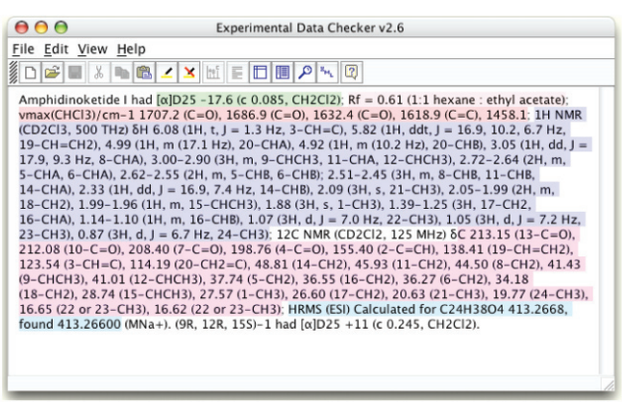

Fig. 7 Amphidinoketide I data with deliberate mistakes.

The most obvious change is the patch of uncoloured text in the middle of the display. The program is not familiar with ^{12}C NMR spectra, and so has not parsed this part of the data. This gap in the colour should make an author or referee look at this part of the data particularly carefully, and help to find the error. The rest of the carbon NMR data has, however, been correctly analysed.

Moving to the *Analysis* view for this incorrect data (Fig. 8) the other errors are highlighted. The left hand side of the window

lists the issues to be checked, in decreasing order of severity. The 'HNMR—no solvent' entry has been selected, and the relevant part of the data is displayed. The program has a list of the usual NMR solvents, and CD_2Cl_3, naturally enough, is not included. The program does not recognise THz as a unit for frequency suitable for an NMR machine, and highlights this. The surplus semicolon has not stopped the program parsing the data correctly.

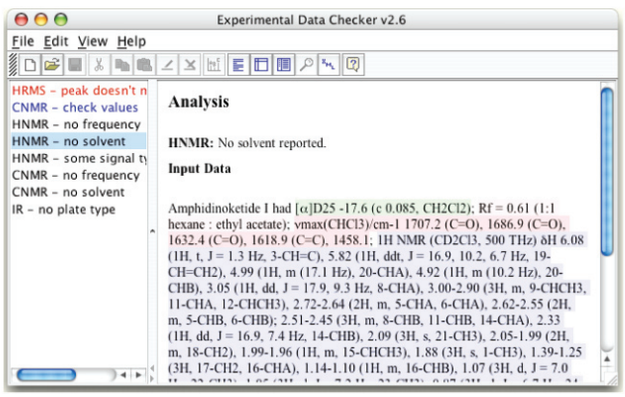

Fig. 8 Analysis of the deliberate mistakes.

The applet and the application are not limited to single molecules. A draft of the experimental part of a PhD thesis (seventy pages) was pasted into the application, and it was parsed in about a minute. One of the options for analysing such a large amount of data is to display it as a table, and this is shown in Fig. 9. It is clear at a glance what data is present and what is missing from the thesis. Clicking on any row of the table gives a detailed analysis of that entry, listing the data which is present and absent. The program attempts to link the chemical name of the molecule to the data, but this is currently a less effective part of the analysis, and requires the compound name to be in the same paragraph as the data. As a result, the table lists some molecules as being unidentified compounds, when their names were given correctly, but the program was unable to find them.

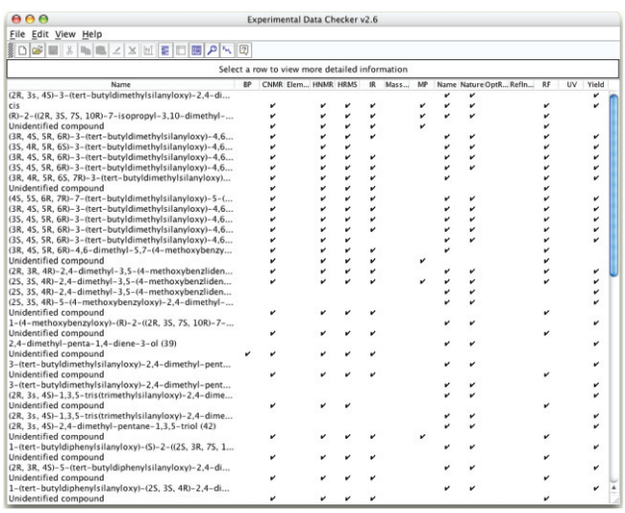

Fig. 9 Table of results.

The toolkit works by using a database of keywords and patterns, based both on the RSC's guidelines for authors (see http://www.rsc.org/is/journals/j1authorinfo.htm) and on two hundred papers which have been published recently in RSC journals. These are used to identify the different components of experimental data, and so discover the name of the molecule, where numbers corresponding to the formula end and where the composition begins, and so on. The toolkit does not assume that authors have followed the RSC's guidelines rigidly, and attempts to analyse whatever information is given. Whilst RSC publications were the focus of development, experimental data from other journals is often parsed correctly by the applet.

Detailed tests were carried out on ten papers that had not been used to develop the keywords. Pasting the whole paper into the applet correctly identified the Experimental section in all the papers, and highlighted 229 paragraphs, of which 172 described compounds. Of these, 153 described single compounds, which were found to contain 815 pieces of experimental data. The applet generated less than 4% false positives, and missed 94 pieces of experimental data (10%). Fourteen of these were due to the authors mis-formatting the data, for example by not opening and closing brackets correctly, and the rest were due to ways of presenting the data which were not wrong but which were non-standard. All of these problems would probably have been caught had the authors or referees had access to this applet and had guidelines to correct the formatting. Non-standard presentation of results would probably be reduced.

Fifty-five cross-correlations are performed to check the data presented. These include:

(i) re-calculating composition from molecular formula

(ii) re-calculating the HRMS ion from molecular formula

(iii) comparing the number of ^{13}C NMR peaks with the number of carbon atoms from HRMS

(iv) comparing the number of hydrogen atoms reported in the ^{1}H NMR data with the number from HRMS

(v) checking spectral peaks appear within appropriate ranges.

A further survey was carried out of a hundred data paragraphs randomly selected from fifty recent papers published in *Organic and Biomolecular Chemistry*, which had not been used in the applet development. These contained a total of 473 items of analytical data, once paragraphs containing information on multiple compounds (4%) had been discarded. The program identified over 92% of these data, and a third of the omissions were due to author formatting errors. The program erroneously identified only fourteen items as data, slightly better than the 4% false positive rate in the first detailed tests.

The program generates four classes of commentary on the data (errors, warnings, comments, information). The applet is sufficiently reliable to be used to identify serious problems that are almost certainly errors. These include problems with mass spectrometry such as a molecular ion with a weight very different to that found from the molecular formula. Thirteen errors were discovered in the survey of the hundred data paragraphs (473 items of data), suggesting a reliability of better than 97%. Warnings are generated when peaks in spectra are outside common values, or when peaks appear to be missing. Thirty-eight warnings were generated. It is likely that most of these issues are correctly reported in the paper, but they are unusual enough for the author to be reminded to consider checking the information. Many comments were also generated, for example, noting when infra-red spectra did not list plate-types, and similar minor issues. Finally, some information is generated, including the number of double-bond equivalents in each molecule, which may be a useful check on the structure.

The applet is small (about 300 kB) and so can be easily accessed, even over a slow internet link, and runs within a web browser. The Java application is more powerful than the applet, and has some functionality that is not available within a web browser (Table 1). For example, it can generate XML, and it can display a whole paper, with the different elements of data highlighted in different colours, so that it is clear at a glance the data that has been parsed and the data that is missing. If the reason for the omission of data is unclear, the website may be useful:‡ http://www.rsc.org/is/journals/checker/run.htm

The program can also be modified readily to become a servlet. This means that it runs on the WWW server, and not on the client machine. All experimental data would be uploaded to the server, so that more complex operations (for example, infra-red spectral

Table 1 Comparison of applet and application

	Applet	Application
Find data	✓	✓
Validate data	✓	✓
Plot spectra	✓	✓
Highlight data	✗	✓
Display tables	✗	✓
Print analysis	✗	✗
Read in text files	✗	✓

simulation, or comparison with large and proprietory databases) could be carried out than are possible on many clients.

With existing published data, searching through reports, theses or libraries is straightforward, if information about a particular molecule is required, but may be difficult if information is needed about compounds with spectral peaks in a particular pattern, or some combination of molecular weight and R_f is an important criterion, or any other requirement to compare the data between different molecules. The problem is much harder if different papers are being compared. When a molecule is reported as being new, it is easy to check if its structure has been previously reported, using SciFinder,[3] Beilstein,[4] or ISIS,[5] but harder to be sure if the spectra reported correspond to spectra which have been reported previously.

These difficulties could be eased if the data were available and stored in a suitably marked-up form. XML provides a method of doing this,[6] but the effort of changing experimental data presented in a traditional way to XML is significant, and would probably not be welcomed by authors. In addition to checking and cross-correlating the experimental data, the application also produces a fine-grained XML version of the paper. This can be stored in databases and analysed using many standard tools, and provides a method of ensuring that the information in the paper can be used to its fullest extent. Since the XML version is stored as a text file, it helps ensure that the data is not lost as the complex standards for word processor documents evolve. The application, therefore, provides a practical method to extract the experimental data from manuscripts and published papers into a structured XML form for storage and analysis.

In conclusion, we have produced a program which should help authors report their experimental data more accurately, help referees check the data more quickly, help sub-editors run automated final checks on manuscripts, help readers produce simulations of spectra, and help research groups and publishers generate marked-up databases of their information. The applications and the documentation are available on: http://www.rsc.org/is/journals/checker/run.htm.

Preliminary tests with this program demonstrate that refereed and published experimental data is highly accurate, but errors are still occasionally perpetuated. The application of the procedure described in this paper will reduce the number of published errors. There are many possibilities for improving the data checking cross-validation, and we welcome suggestions. We hope that it will be widely used, and any issues or parsing errors that arise will be communicated to the authors so that future versions of the program may be even more effective in assisting scientists in the production and analysis of precise and clear papers.

Notes and references

† Amphidinoketide I had $[\alpha]_D^{25}$ −17.6 (*c* 0.085, CH_2Cl_2); $\mathbf{R_f}$ = 0.61 (1:1 hexane:ethyl acetate); $\nu_{max}(CHCl_3)/cm^{-1}$ 1707.2 (C=O), 1686.9 (C=O), 1632.4 (C=O), 1618.9 (C=C), 1458.1; **^{1}H NMR** (CD_2Cl_2, 500 MHz) δ_H 6.08 (1H, t, *J* = 1.3 Hz, 3-C*H*=C), 5.82 (1H, ddt, *J* = 16.9, 10.2, 6.7 Hz, 19-C*H*=CH_2), 4.99 (1H, m (17.1 Hz), 20-CH_A), 4.92 (1H, m (10.2 Hz), 20-CH_B), 3.05 (1H, dd, *J* = 17.9, 9.3 Hz, 8-CH_A), 3.00–2.90 (3H, m, 9-C*H*CH_3, 11-CH_A, 12-C*H*CH_3), 2.72–2.64 (2H, m, 5-CH_A, 6-CH_A), 2.62–2.55 (2H, m, 5-CH_B, 6-CH_B), 2.51–2.45 (3H, m, 8-CH_B, 11-CH_B, 14-CH_A), 2.33 (1H, dd, *J* = 16.9, 7.4 Hz, 14-CH_B), 2.09 (3H, s, 21-CH_3), 2.05–1.99 (2H, m, 18-CH_2), 1.99–1.96 (1H, m, 15-C*H*CH_3), 1.88 (3H, s, 1-CH_3), 1.39–1.25 (3H, 17-CH_2, 16-CH_A), 1.14–1.10 (1H, m, 16-CH_B), 1.07 (3H, d, *J* = 7.0 Hz, 22-CH_3), 1.05 (3H, d, *J* = 7.2 Hz, 23-CH_3), 0.87 (3H, d, *J* = 6.7 Hz, 24-CH_3); **^{13}C NMR** (CD_2Cl_2, 125 MHz) δ_C 213.15 (13-*C*=O), 212.08 (10-*C*=O), 208.40 (7-*C*=O), 198.76 (4-*C*=O), 155.40 (2-*C*=CH), 138.41 (19-*C*H=CH_2), 123.54 (3-*C*H=C), 114.19 (20-*C*H_2=C), 48.81 (14-*C*H_2), 45.93 (11-*C*H_2), 44.50 (8-*C*H_2), 41.43 (9-*C*HCH_3), 41.01 (12-*C*HCH_3), 37.74 (5-*C*H_2), 36.55 (16-*C*H_2), 36.27 (6-*C*H_2), 34.18 (18-*C*H_2), 28.74 (15-*C*HCH_3), 27.57 (1-*C*H_3), 26.60 (17-*C*H_2), 20.63 (21-*C*H_3), 19.77 (24-*C*H_3), 16.65 (22 or 23-*C*H_3), 16.62 (22 or 23-*C*H_3); **HRMS** (ESI) Calculated for $C_{24}H_{38}O_4$ 413.2668, found 413.26600 (MNa^+). **(9*R*, 12*R*, 15*S*)-1** had $[\alpha]_D^{25}$ +11 (*c* 0.245, CH_2Cl_2).

‡ Problems may include: (i) the program freezes when text is pasted in: this may be a result of pasting too much at one time. Either cut and paste in smaller sections, or else save the data file as a 'text-only' file, and open this file from the application. (ii) the text looks reasonable, but is not parsed: this may be due to the presence of special characters. For example, the program can usually handle Greek letters, but these may be presented in a format it cannot understand. A possible solution is to save the data file as a 'text-only' file, and try again.

1 J. M. Goodman and L. M. Walsh, *Chem. Commun.*, 2003, 2616.
2 Open source artistic license: http://www.opensource.org/licenses/artistic-license.php.
3 SciFinder: Chemical Abstracts Service, 2540 Olentangy River Road, PO Box 3012, Columbus, OH 43210, USA.
4 Beilstein: MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA 94577, USA.
5 ISIS: MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA 94577, USA.
6 P. Murray-Rust, H. S. Rzepa, M. Wright and S. Zara, *Chem. Commun.*, 2000, 1471.