An Introduction to readJDX

Bryan A. Hanson[a]

[a] Dept. of Chemistry & Biochemistry, DePauw University; hanson@depauw.edu

This version was compiled on October 15, 2018

This vignette is based on `readJDX` version 0.3.250.

Background

The JCAMP-DX format was developed as an manufacturer-independent means of sharing spectroscopic data (www.jcamp-dx.org). The standard is described in a series of publications (McDonald and Wilks (1988); Grasselli (1991); Davies and Lampen (1993); Lampen *et al.* (1994, 1999); Baumbach *et al.* (2001); Cammack *et al.* (2006); Woollett *et al.* (2012)). JCAMP-DX was developed during a time when data storage was expensive, and hence makes extensive use of compression schemes. The original application was to IR spectroscopy, but the standard has evolved over time to accommodate other spectroscopies.

File Structure

JCAMP-DX files consist of two parts:

- A more-or-less human readible set of metadata which is needed to understand the data and carry out the decompression. Besides required basic information about the data itself, most files contain instrument and manufacturer-specific parameters in the metadata.
- A variable list, compressed in various ways.

Challenges When Reading Files

The JCAMP-DX standard allows a lot of flexibility and instrument manufacturers have written widely varying export functions. Some of the challenges in reading a JCAMP-DX file include:

- JCAMP-DX files can contain different kinds of data, including non-spectroscopic data (Gasteiger *et al.* (1991)) and more than one type of spectroscopic data.
- JCAMP-DX files can contain more than one spectrum in the file.
- Instruments may be configured to use . or , as the decimal point when writing files. This is generally a geographical / cultural nuance.
- Numbers may be written using E to signify exponent, but only in some compression formats.
- The variable list can be presented in several possible formats.
- Some manufacturers take liberties with the required format.

Supported Formats

- Variable lists can be presented in several different formats. The supported formats are:

  - XYDATA=(X++(Y..Y)) Each line starts with an x value, and is followed by as many y values as can fit within the 80 character per line limit. Subsequent x values are incremented according to the x resolution and the number of y values that fit on the previous line (which in turn depends upon the compression scheme).
  - DATA TABLE=(X++(R..R)) As above. The real data from a 1D NMR spectrum.
  - DATA TABLE=(X++(I..I)) As above. The imaginary data from a 1D NMR spectrum.
  - DATA TABLE=(F2++(Y..Y)) As above. Format used for the slices of a 2D NMR spectrum.

- Within a variable list, several different compression schemes can be employed. The following are supported:

  - AFFN: ASCII numbers separated by at least one space.
  - PAC: Numbers separated by exactly one space, + or -.
  - SQZ: Delimiter, leading digit and sign are replaced by a pseudo-digit. A pseudo-digit is typically a letter.
  - DIF: DIF uses a SQZ pseudo-digit for the first y value, but subsequent y entries are differences between each data value after the first. Sometimes referred to as SQZDIF.
  - DUP: Not a format, but a method of signifying repeated values.
  - DIFDUP: A combination of DIF and DUP. Widely used, as it permits the greatest amount of compression.

Formats That are Not Supported

- Mixed spectroscopic types and non-spectroscopic entries (such as structures) are not supported by `readJDX` and will not be supported in the future.
- Compound files: JCAMP-DX files may contain more than one spectrum in the file. These types of files are not currently supported (however, 1D and 2D NMR data are supported, even though they contain more than one spectrum). The following JCAMP-DX standards require a compound file and are therefore not supported:

- EMR, EPR, ESR spectroscopy (Cammack *et al.* (2006))
- CD spectroscopy (Woollett *et al.* (2012))

- Variable lists of simple x, y pairs (format XY..XY) are not supported. Your pull requests to deal with this format are welcomed.
- `readJDX` is geared toward raw spectral data. Therefore variable lists formats representing derived information like PEAK TABLE and PEAK ASSIGNMENTS are not supported (but again, your pull requests are welcomed!).

Practical Matters

`readJDX` tries its best to deal with all these options. If you have a file that you believe should be supported but gives an error, please file an issue at GitHub. Be sure to attach the file that is giving you problems.

Before release, `readJDX` is tested against a large collection of files with varying formats. A few of these files were obtained locally. Others were collected from publically available sources (e.g. http://www.jcamp-dx.org/testdata.html). These files are not included with the package to save space, and in addition, while they are publically available, for many of them the licensing status is unclear (i.e. the OWNER entry).

The JCAMP standard requires a number of checks on the integrity of the data decompression process. `readJDX` implements most of these either directly or indirectly. Verification is important, and we have found JCAMP files that were not written correctly in the process of checking integrity. For details about how data decompression is checked, please see the original source files.

Known Issues

- Some test files in the author's possession fail when carrying out the y value check, for unknown reasons.

Program Flow

`readJDX` is coded in such a way that it should be easy to add features. Contributions to improve or expand the package, including pull requests, are always welcome! The scheme below shows how control of the program is organized. The $\rightarrow$ means "calls" and $\overset{?}{\rightarrow}$ means "may call".

```
readJDX
→ findDataTables
→ extractParams
→ processDataTable
    →? decompressJDXxyy
        → getJDXcompression
    →? insertDUPs
        → repDUPs
    →? unSQZ
    →? deDIF
        → unDIF
→? functions to decompress other formats (future?)
```

References

Baumbach J, Davies A, Lampen P, Schmidt H (2001). "JCAMP-DX. A Standard Format for the Exchange of Ion Mobility Spectrometry Data - (IUPAC recommendations 2001)." *Pure and Applied Chemistry*, **73**(11), 1765–1782. ISSN 0033-4545. .

Cammack R, Fann Y, Lancashire R, Maher J, McIntyre P, Morse R (2006). "JCAMP-DX for electron magnetic resonance(EMR)." *Pure and Applied Chemistry*, **78**(3), 613–631. ISSN 0033-4545. .

Davies A, Lampen P (1993). "JCAMP-DX for NMR." *Applied Spectroscopy*, **47**(8), 1093–1099. ISSN 0003-7028. .

Gasteiger J, Hendricks BMP, P H, C J, H S (1991). "JCAMP-CS: A Standard Exchange Format for Chemical Structure Information in a Computer-Readable Form." *Applied Spectroscopy*, **45**(1), 4–11.

Grasselli J (1991). "JCAMP-DX, A Standard Format for Exchange of Infrared-Spectra in Computer Readible Form." *Pure and Applied Chemistry*, **63**(12), 1781–1792. ISSN 0033-4545. .

Lampen P, Hillig H, Davies A, Linscheid M (1994). "JCAMP-DX for Mass Spectrometry." *Applied Spectroscopy*, **48**(12), 1545–1552. ISSN 0003-7028.

Lampen P, Lambert J, Lancashire R, McDonald R, McIntyre P, Rutledge D, Frohlich T, Davies A (1999). "An Extension to the JCAMP-DX Standard File Format, JCAMP-DX V.5.01 (IUPAC Recommendations 1999)." *Pure and Applied Chemistry*, **71**(8), 1549–1556. ISSN 0033-4545. .

McDonald R, Wilks P (1988). "JCAMP-DX, A Standard Format for Exchange of Infrared-Spectra in Computer Readible Form." *Applied Spectroscopy*, **42**(1), 151–162. ISSN 0003-7028. .

Woollett B, Klose D, Cammack R, Janes RW, Wallace BA (2012). "JCAMP-DX for circular dichroism spectra and metadata (IUPAC Recommendations 2012)." *Pure and Applied Chemistry*, **84**(10), 2171–2182. ISSN 0033-4545. .