Journal of Physics: Conference Series **190** (2009) 012033 doi:10.1088/1742-6596/190/1/012033

Towards a black-box for biological EXAFS data analysis – III. A universal post-processor for fluorescence XAS: KEMP2

Gerd Wellenreuther[1], Wolfram Meyer-Klaucke[2]

[1] HASYLAB at DESY, Notkestraße 85, 22637 Hamburg, Germany

[2] EMBL, Outstation Hamburg, Notkestraße 85, 22603 Hamburg, Germany

Gerd.Wellenreuther@desy.de

Abstract. Because of the typically rather low trace element concentrations in biological samples XAS is collected in fluorescence mode with the help of a multi-element detector. In addition, several scans have to be accumulated and thus post-processing operations have to be carried out after the measurement, e.g. selecting good detector channels, checking for radiation damage, fitting of background and normalization. KEMP2 is a freely available post-processor implementing all these steps, while maintaining universal applicability by separating the instrument- and beamline-dependent steps like energy-calibration etc. from XAS-generic tasks. Thus, KEMP2 has the potential to serve as a highly automated data pipeline assisting both senior and novice users.

1. Introduction

Trace elements are essential for the function of biological systems. They are involved in central biological processes such as respiration, metabolism, photosynthesis, cell division, muscle contraction, nerve impulse transmission, and gene regulation. Metal ions, for example, stabilize proteins, regulate of gene expression, transport electrons and oxygen, induce protein activity and form active sites of enzymes. X-ray absorption spectroscopy (XAS) is the only element-specific technique providing electronic and geometric structural information on trace elements in biological samples. Thus, XAS can help to understand their chemistry, to identify binding motifs and thereby lead to a prediction of potential protein functions.

Since even the most concentrated protein samples are dilute, XAS measurements are almost exclusively performed in fluorescence mode, typically using multi-element detectors. Still, several XAS-scans have to be accumulated to obtain data required to extract a good EXAFS-signal with a sufficient k-range [1].

Unfortunately, it is not uncommon that several detector elements are suffering from artifacts, e.g. due to ice-crystal formation in the sample causing reflections at certain energies. Another class of common artifacts is glitches caused by multi-beam diffraction in the monochromator. During post-processing either these artifacts or the entire scans have to be removed from the final data pool. Further steps required include the fitting of background, extracting the XANES and EXAFS as well as calculating a first FT.

Despite the fact that most of these operations have to be carried out for biological XAS data collected around the world, no “silver bullet” like a universal post-processor for fluorescence XAS

data exists to date. This is mainly caused by the varying instrumentation and data-formats used, which is in strong contrast to the situation in macromolecular X-ray crystallography (MX), where the entire pipeline from crystallization trials through data acquisition until the final modeling is achieved by a strongly automated pipeline. KEMP2 attempts to overcome this restriction by separating the few instrument and data-format dependent tasks from the core of the generic XAS post-processing tasks. It is a further development of the post-processing script KEMP[2], which was for years used at beamline D2 (DORIS, Hamburg, operated by EMBL Hamburg outstation)[3-10].

Such a generic post-processor is a necessary step toward automated treatment of BioXAS data yielding the proper model and publication-ready graphs from the raw XAS-data. With the algorithm ABRA[11-13] it was already shown that the automatic refinement and analysis of BioXAS data is feasible, but this step requires a properly averaged and background-subtracted dataset. Using KEMP2 any user at any beamline should be able to accurately extract the EXAFS-signal from his biological fluorescence data, to be either used as an input for ABRA or for any other XAS-modeling software, e.g. IFEFFIT[14], WinXAS[15], GNXAS [16] or EXCURV [17] .

2. Implementation

In this part, the functionality of KEMP2 is sketched. Details of the implementation are given for the most crucial steps of the post-processing of XAS-data. The decision to use the scripting language Python for KEMP2 was a deliberate one, since Python is free, easy to use, and its powerful libraries for scientific programming / plotting (e.g. SciPy+NumPy, Matplotlib) aides quick development. Consequently, Python is typically available at synchrotrons around the globe. KEMP2 itself should be easy to install and use by any user, requiring very little knowledge of XAS or computing, while giving experts the tools to quickly manipulate raw data taken at any instrument. It requires a small number of freely available modules (SciPy+NumPy , Matplotlib, AstroAsciiData, mpfit), which are easy to install even for inexperienced users. The installation is also described on KEMP2's homepage [18].

In order to be able to use raw data from virtually every instrument the post-processing has to be divided in two parts: First, in the **instrument and data-format specific part**, the raw data has to be converted into the standard representation used within KEMP2's post-processing part. All scans in a given directory will be processed, which is also where KEMP2 will direct its output.

2.1. Raw-data conversion (RDC)

KEMP2's standard representation is a comma-separated-value (CSV) file, typically using semicolons to separate values, and any line starting with a hash ('#') is regarded as a comment. The first column contains the energy in eV, the second column contains the signal from the I0-monitor already corrected for offset. Any further columns are interpreted as fluorescence detector channels. Since KEMP2 later has to sum several data sets from different scans it requires that the energy axis in the first column is always the same.

Consequently, this defines what has to be done for any beamline in the instrument-dependent step:

1. Read-in the instrument-specific data
2. Subtract offset from I0-signal
3. Correct/calibrate energy axis in eV
4. Interpolate the monitor/fluorescence signals on generic energy axis
5. Write generic KEMP2-files

It varies from instrument to instrument how much effort has to be put into these individual steps, e.g. a proper monochromator with encoders can render step 3. and/or 4. unnecessary. A few such routines have already been implemented, and are available along with KEMP2. Otherwise, it should be a small effort to re-write the existing routines.

2.2. Post-processing

Currently, KEMP2 is still developing, but it already includes the most vital features for XAS data post-processing. These are currently:

1. Normalize detector elements using monitor column
2. Summing over the number of scans for individual detector elements
3. Selecting the "good" detector channels (at present interactive, will be automated soon)
4. Summing over the number of selected detector channels
5. Displaying the time-development of XANES
6. Background-subtraction / deglitching (iterated until user is satisfied):
    a. Subtract the pre-edge background by fitting a constant
    b. Fit a spline to the data (interactive)
    c. Normalize the edge jump / Subtracting spline
    d. Extract / Display XANES & EXAFS
    e. Calculate / Displaying Fourier-Transform
    f. Remove glitches (interactive)
7. Outputting of XANES & EXAFS, FT etc.

2.2.1. *Selecting the "good" detector elements.* A plot of all detector elements is shown, and the user is asked whether he would like to use each detector element. E.g. detector elements showing extended artifacts from ice crystal formation in the sample should be discarded, while elements containing only glitches from the monochromator can be used and deglitched later.

2.2.2. *Displaying the time-development of XANES.* It is good scientific practice to check whether radiation damage occurred during data collection. Therefore, KEMP2 plots the normalized data averaged only over the selected number of detector elements. Consequently, every scan gives a single curve, and all curves should lie on top of each other, except for statistical variations. Especially a drop of the white line intensity would be associated with radiation damage. Currently, no actions are implemented by KEMP2, e.g. no automatic exclusion of damaged scans – if radiation damage is occurring the user is required to remove the corresponding files from the working directory and start again.

2.2.3. *Background / de-glitching.* Several parameters are required for a proper removal of the background and later during data extraction (edge position, region of pre-edge background, regions for XANES / EXAFS-extraction). These can be supplied by copying an appropriate old KEMP-style "remove.par"-file into the working directory, but this is not mandatory. In any case, the user is given the opportunity to change these values.

A constant background is fitted to the pre-edge background region and subtracted from the data. The user then can choose the number of inner knots of the spline (equidistant in energy) used to fit the slowly varying background above the edge. The spline is fitted and used to normalize the XANES-spectrum as well as to subtract the background from the EXAFS-region. In order to aid the user in achieving a proper fit the extracted XANES in energy space and the k-weighted EXAFS in k-space as well as the Fourier transform (without any kind of phase correction) are displayed immediately. Since the results of this process are immediately displayed in energy-, k- and R-space together with previous results the user can optimize his spline simultaneously in all regimes.

In addition, the user can deglitch his data in a very easy fashion: First, he is required to enter the approximate k-value of a glitch. Now, the very data points at this k-value are plotted over their index number. The use can successively enter any number of glitch-points which will then be excluded from the data. This whole procedure can be iterated until the user is happy with the achieved normalization/background subtraction/deglitching.

2.2.4. *Outputting of XANES & EXAFS, FT etc.* Finally, KEMP2 produces several, blank-separated-files for later use in ABRA or for the generation of publication-ready graphs:

- The raw, unnormalized but **averaged data** over E in eV
- The extracted, **normalized XANES** over E in eV

- The **absorption** over E above the edge in eV
- The **k-weighted EXAFS** above the edge over k in $Å^{-1}$
- The **magnitude of the uncorrected Fourier** transform over R in Å

In addition, the position of the glitches is written in "KEMP2_glitches.log" – giving the user the exact k-value and energy above the edge of the points removed from the averaged data.

3. Discussion/Conclusions/Outlook

KEMP2 was successfully used to process data from the beamline DUBBLE at the ESRF, Grenoble, France, and from beamline SuperXAS at the SLS, Villigen, Switzerland. Consequently, the necessary RDC-scripts for both instruments exist. DUBBLE is using the data format developed at the Daresbury Laboratory, so for any other beamline using this data-format only instrument-dependent differences have to be implemented in the RDC-step. Currently, the functionalities provided by KEMP2 are still being improved, future developments will include (i) interactive/automatic removal of scans suffering from radiation damage and (ii) automatic quality control / suggestion of choice of "good" detector channels (as implemented in KEMP[19])

On the long term, we expect KEMP2 to play a central role in data reduction for biological XAS because it both assists the novice users and makes the work of experts very efficient due to its high degree of automation.

References

[1] Ascone I, Meyer-Klaucke W and Murphy L 2003 *J Synchrotron Radiat* **10** 16-22
[2] Korbas M, Marsa D F and Meyer-Klaucke W 2006 *Rev Sci Instrum* **77** -
[3] Peroza E A, Al Kaabi A, Meyer-Klaucke W, Wellenreuther G and Freisinger E 2009 *J Inorg Biochem* **103** 342-53
[4] Hollenstein K, Comellas-Bigler M, Bevers L E, Feiters M C, Meyer-Klaucke W, Hagedoorn P L and Locher K P 2009 *J Biol Inorg Chem* **14** 663-72
[5] Wellenreuther G, Cianci M, Tucoulou R, Meyer-Klaucke W and Haase H 2009 *Biochem Bioph Res Co* **380** 198-203
[6] Hiromoto T, Ataka K, Pilak O, Vogt S, Salomone-Stagni M, Meyer-Klaucke W, Warkentin E, Thauer R K, Shima S and Ermler U 2009 *Febs Lett* **583** 585-90
[7] Todorovic S, Justino M C, Wellenreuther G, Hildebrandt P, Murgida D H, Meyer-Klaucke W and Saraiva L M 2008 *J Biol Inorg Chem* **13** 765-70
[8] Toussaint L, Cuypers M G, Bertrand L, Hue L, Romao C V, Saraiva L M, Teixeira M, Meyer-Klaucke W, Feiters M C and Crichton R R 2009 *J Biol Inorg Chem* **14** 35-49
[9] Minicozzi V, Stellato F, Comai M, Serra M D, Potrich C, Meyer-Klaucke W and Morante S 2008 *J Biol Chem* **283** 10784-92
[10] Shima S, Pilak O, Vogt S, Schick M, Stagni M S, Meyer-Klaucke W, Warkentin E, Thauer R K and Ermler U 2008 *Science* **321** 572-5
[11] Wellenreuther G and Meyer-Klaucke W 2007 *AIP Conference Proceedings* 322-32
[12] Wellenreuther G and Meyer-Klaucke W 2009 *submitted.*
[13] Wellenreuther G and Meyer-Klaucke W 2007 http://webapps.embl-hamburg.de/exafs/exafs_new.html
[14] Newville M 2001 *J Synchrotron Radiat* **8** 322-4
[15] Ressler T 1998 *J Synchrotron Radiat* **5** 118-22
[16] Westre T E, Dicicco A, Filipponi A, Natoli C R, Hedman B, Solomon E I and Hodgson K O 1995 *J Am Chem Soc* **117** 1566-83
[17] Binsted N, Strange R W and Hasnain S S 1992 *Biochemistry* **31** 12117-25
[18] Wellenreuther G 2009 http://www.embl-hamburg.de/~gwellenr/KEMP2/KEMP2.htm
[19] Lippold B, Meyer-Klaucke W, Meyer T and Henkel G 2005 T *J Synchrotron Radiat* **12** 45-52