Journal of
Synchrotron
Radiation

ISSN 0909-0495

Received 14 September 2003
Accepted 5 November 2004

Towards an automated quality control of XAS data

Björn Lippold,[a] Wolfram Meyer-Klaucke,[b] Thomas Meyer[a] and Gerald Henkel[c]*

[a]Universität Duisburg-Essen, Standort Duisburg, Fakultät für Naturwissenschaften, Institut für Chemie – Festkörperchemie, 47048 Duisburg, Germany, [b]EMBL Outstation Hamburg, c/o DESY, Notkestrasse 85, 22607 Hamburg, Germany, and [c]Universität Paderborn, Chemie und Chemietechnik, Warburger Strasse 100, 33098 Paderborn, Germany.
E-mail: biohenkel@fb13n.uni-paderborn.de

The quality of fluorescence X-ray absorption spectroscopy (XAS) data strongly depends on the identification and elimination of contributions suffering from artificial deviations. To enhance detection of deviations, XAS data are converted here to difference spectra and cumulative difference spectra. A variety of statistical criteria and procedures are examined for their application in the quality control of such data. The criterion best suited in this case is determined and a strategy for the automatic elimination of artefacts is developed: deviation-affected spectra are iteratively removed from the data pool. A threshold is defined to avoid unnecessary reduction of the experimental data pool. Exemplarily the procedure is applied for the quality control of BioXAS data.

Keywords: X-ray absorption spectroscopy (XAS); EXAFS; quality control; biological system; BioXAS; metalloproteins.

1. Introduction

X-ray absorption spectroscopy (XAS) provides an excellent probe for electronic or geometric structure determination of metal sites in biological molecules. Since current genomic projects produce and isolate an increasing number of interesting biological compounds, both natural and synthetical, XAS begins to face the challenge of handling and analyzing huge quantities of complex data (Ascone *et al.*, 2003). Simultaneously an increasing interest in the small differences between metal sites or different protein states requires a more sophisticated analysis of the data. Hence, a careful and efficient quality control of the collected data becomes a vital step for advanced high-throughput XAS projects.

For BioXAS experiments and other spectroscopic techniques, several scans are accumulated and averaged for final data evaluation (Ranieri-Raggi *et al.*, 2003). With each accumulated spectrum the noise level of the averaged data decreases and thus the information content can improve. While slight statistical variations for each individual spectrum are expected, there are in some cases non-statistical deviations which negatively affect the quality of the final data set. Thus, quality control can be based on the elimination of individual spectra suffering from significant deviations from the average of the remaining data pool. In this way, data quality is assured by limiting the data quantity. Obviously, all deviant spectra must be eliminated from the data pool. At the same time, as many spectra as possible must be retained to maintain a good signal-to-noise ratio. Therefore, balancing both quality and quantity of the data is the key to efficient quality control.

Biological samples for XAS measurements are typically available as solutions in low concentration resulting in a comparatively high noise level for each individual scan. Therefore a high number of accumulated scans is required. Usually such samples are measured in fluorescence mode by multi-element detectors (Ascone *et al.*, 2003). For each of the corresponding detector channels one spectrum is measured per scan so the final XAS data pool consists of an $n \times m$ matrix of accumulated individual contributions (n: number of scans; m: number of detector elements/channels).

While spectra of highly concentrated compounds feature a low noise level that allows visible data screening, even for experienced researchers BioXAS data quality is more difficult to assess in the case of low metal concentrations. Objective criteria for quality control of XAS data should help to replace experience-based subjective criteria. For reliable quality control the human influence on the procedure has to be minimized. This requires partial or, even better, full automation. Moreover, such a system can be linked directly to automated data-collection systems which are becoming available at more and more BioXAS stations.

Diverse mathematical functions can be used for qualitative detection and quantitative analysis of individual spectra contributing to the data pool. Criteria should be chosen such that they can easily be included in software routines of typical XAS data-processing software and applications. This minimizes the human influence on the quality control and enables even inexperienced users to use BioXAS methods.

If applied during the data-collection process as quality monitoring, the quality control systems help to optimize the

required data quantity, resulting in a more efficient beam-time usage. Moreover, a detailed online analysis and identification of the type of deviations can probably also help to identify immediately the presence of potential artefacts, like sample inhomogenities or spectrometer malfunctions. Appropriate protocols based on such objective criteria might result in the development of either automatic correction procedures or guidelines for quality improvement.

In this publication, several statistical criteria for quality control are compared. Additionally, mechanisms for the differentiation between significant deviations and acceptable statistical variations are discussed, resulting in the first functional quality control procedure.

2. Quality control

2.1. Types of deviations in spectra

Spectroscopic data are typically influenced by a certain noise level which causes statistical variations in the individual data points. Besides these statistical variations (Fig. 1*a*), there are several typical deviations which can affect recorded spectra (see Figs. 1*b*–1*e*). The most common and least obvious is the shift of a small group of data points resulting in a vertical offset (Fig. 1*b*). Related but more serious deviations are jump discontinuities or saltuses which emerge when, from a certain energy onwards, the spectrum is shifted along the *y* axis (Fig. 1*c*). Sudden artificial changes of the slope at a certain energy are also common deviations (Fig. 1*d*). Additionally, the spectrum may be convoluted by some kind of oscillating function causing a periodic bending of the spectrum (Fig. 1*e*). Obviously, in experimental data, combinations of the different deviations might exist.

2.2. Experimental data pool

XAS data gathered by multi-channel detectors with m detector elements are stored as m individual spectra contributions per scan. It can be expected that deviations occur either during a certain time frame and affect an entire scan or manifest in all contributions of a specific channel if they are caused by electronic effects or sample preparation/mounting (*e.g.* inhomogeneous samples, icing, glitches).

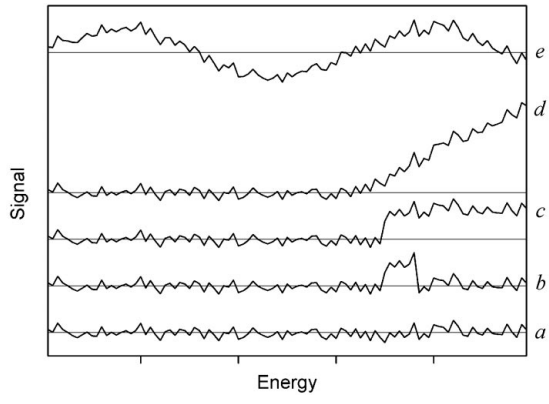

Figure 1
Examples of variations and deviations from a flat-line spectrum: (*a*) statistical variations, (*b*) shift of group of data points, (*c*) jump discontinuity, (*d*) change of slope, (*e*) oscillation.

Depending on the number of scans and detector channels, it is sometimes useful to analyze specific subsets of the data pool instead of single contributions. In turn, subsets based on n scans and m channels can be analyzed instead of the $n \times m$ individual contributions. Thereby the quality control procedure is simplified and potential scan- or channel-dependent deviations become more obvious. For the complete quality control procedure both scan- and channel-based subsets have to be checked successively.

2.3. Difference spectra

Evaluation of the quality of measured spectra requires a reference. Typically, theoretical references are unavailable prior to the data collection. Therefore, they can be calculated from the data pool containing all measured spectra. This assumes that at least the majority of contributions are free from artefacts. Then the reference for each subset is defined by the average over all other normalized subsets. If the difference between a non-deviating subset and its reference is plotted, approximately a flat line is expected. Ideally, variations should only correspond to the actual noise level. Any significant discrepancy from the expected flat line indicates a difference between subset and reference. This indicates that the subset features non-statistical deviations and should probably be excluded from the data pool.

2.4. Cumulative difference spectra

In many cases the noise level of the difference spectra is very high. Hence, it might be difficult to detect possible deviations. While the noise cancels out by averaging all contributions, the hidden deviations remain and affect the data quality negatively.

One way to detect such hidden deviations is to sum up the data points of the difference spectrum $D_{(i)}$ successively,

$$A_{(j)} = \sum_{i=1}^{j} D_{(i)}. \qquad (1)$$

In the absence of artefacts the resulting *cumulative spectrum* $A_{(j)}$ is expected to result in a flat line, indicating that all statistical variations cancel out. In the case of deviations the differences are summed up and thereby become more evident. Thus the introduction of cumulative difference spectra should increase the significance and facilitate the detection of deviations.

To study this effect a statistical distribution for 99 data points was simulated and the four aforementioned typical deviations were introduced (see Fig. 1). The cumulative difference spectra derived from these data illustrate the effect (Fig. 2). If a spectrum exhibits only noise the procedure yields a flat line (Fig. 2*a*), whereas the discrepancies in the other spectra are amplified (Figs. 2*b*–2*e*). This effect can be quantified by suitable statistical criteria (Table 1). The comparison of standard deviations from linear regression (see below) shows significant differences for the analysis of difference and

Table 1
Standard deviations of calculated variation- and deviation-affected flat-line spectra and their respective cumulative difference spectra.

	Calculated flat-line spectra (Fig. 1)	Cumulative difference spectra (Fig. 2)
(*a*) Statistical variations	0.97	2.82
(*b*) Shift of group of data points	1.52	9.69
(*c*) Jump discontinuity	2.09	35.60
(*d*) Change of slope	3.61	59.94
(*e*) Oscillation	3.62	37.43

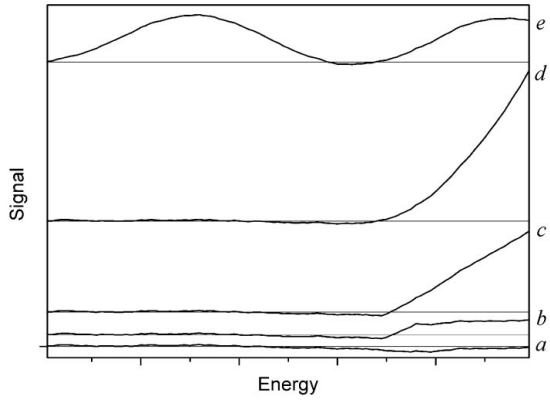

Figure 2
Cumulative difference spectra of variation- and deviation-affected flat-line spectra: (*a*) statistical variations, (*b*) shift of group of data points, (*c*) jump discontinuity, (*d*) change of slope, (*e*) oscillation.

cumulative difference spectra. While the exemplary shift of seven data points increases the standard deviation of the calculated spectrum (Fig. 1*b*) by merely 57%, the increase for the cumulative spectrum (Fig. 2*b*) is 244%. This effect is even more pronounced for the other types of artefacts, which considerably simplifies their detection.

2.5. Iterative procedure

Ideal for the identification of all deviant subsets in the data pool is an iterative procedure. If the data pool contains a subset with significant deviations, the references of all other subsets are affected. If the influence of the deviant subset is significant enough, this may negatively affect the comparisons between the other subsets and their references. But even in this case the deviation from the reference is most pronounced for the non-standard subset allowing its identification without doubt. After excluding this contribution from the data pool, the quality control procedure is restarted. Now either all remaining subsets are of similar quality or again the one with the largest discrepancy can be identified. This procedure is repeated and subsets starting from the worst to the best are removed until the desired quality level is reached.

2.6. Criteria

For the actual quality control process a suitable criterion or a set of criteria is required. Based on the fact that for both the difference and cumulative difference spectra flat lines with only small variations from the noise are expected, several well established criteria are readily available.

An ideal flat-line difference spectrum can be either interpreted as a group of data points which is spread statistically around the expected zero value or as a constant function $[f(x) = 0]$ which can be analyzed by means of linear regression. Difference spectra are suited for the first approach while for cumulative difference spectra the second approach can be applied.

Several criteria can be used for investigation of the difference spectra. For the data-point-based approach the largest positive or negative deviation from the ideal flat line (criterion 1) or the sum of all positive, negative (criterion 2) or the complete set of deviations (criterion 3) can be employed. Additionally, criterion 4 calculates the deviation of the set of data points from its mean value. The kurtosis criterion (criterion 5) provides insights into the shape of a data-point distribution and therefore distortions from the ideal flat line (Abramowitz & Stegun, 1972).

Instead of analyzing the expected flat line as a group of independent data points, criteria from linear regression are available for the cumulative difference spectra (Edwards, 1976). The overall quality of an approximated linear function can be determined by the correlation coefficient (criterion 6) or the standard deviation (criterion 7). As the slope of the linear function (criterion 8) is expected to be zero, its determination by linear regression is also suited.

All aforementioned criteria are tested for the purpose of quality control. They are compared with the visual plots of the difference spectra and the significance of detected deviations and variations is verified for experimental data (see below). The different criteria and their suitability for quality control procedures are summarized in Table 2.

Of all criteria tested only the correlation coefficient (6) does not indicate real deviations reliably. For the identification of certain deviations the kurtosis (5) and slope (8) criteria are sometimes useful and both can be used for supplemental information if required. Sometimes additional results are achieved through careful combination of criteria 1, 2 and 3. It seems that these criteria allow a more detailed diagnosis of the specific type of deviation. In future, a more complex analysis of the difference spectra using all supplemental criteria might help to identify the type of deviation, automatically allowing immediate adaptation and improvement of the experimental conditions.

The mean deviation (4) for difference spectra and the standard deviation from linear regression (7) for cumulative difference spectra are the best candidates for quality control. Criterion 4 is successfully applied to most of the data sets tested but in a few cases it identifies spectra with a high noise level instead of ones with significant deviations. Even in these cases the results from criterion 7 correspond to the visual impression of the spectra (Fig. 3). The validity of the selection of subsets based on criterion 7 could be proven for all tested experimental data sets. While supplemental criteria allow a more detailed investigation on the nature of the deviations,

Table 2
Summary of statistical criteria usable for quality control of spectroscopic data.

Criterion	Description	Results for the test cases
1. Largest positive/negative deviation (difference spectra)	The largest positive or negative value for the deviation of one data point is determined.	These criteria indicate only a singular deviation. Although the extreme deviation of a single data point often corresponds to a larger deviation, in some cases it is statistically irrelevant.
2. Summed positive/negative deviations (difference spectra)	All positive or negative deviations of the difference spectrum from the flat line are summed up.	High values for each of both criteria often signify strong deviations, especially if one of the criteria is significantly higher than the other (see criterion 3). If both values are high but of similar order of magnitude, these criteria sometimes simply indicate a high noise level.
3. Summed deviations (difference spectra)	For this criterion all deviations of the difference spectrum from the flat line are summed up. For statistical variations a value of approximately zero is expected.	In many cases this criterion correlates to the efficient criteria and can be used with criterion 2 for a more detailed interpretation of variations and deviations. Although a high value of this criterion typically indicates deviations, a low value may correspond to the distribution of deviations.
4. Mean deviation (difference spectra)	This criterion calculates the overall absolute deviation of all data points from their mean value.	The mean deviation is a useful criterion for quality control of spectra. Although it does not take into account the shape of difference spectra, the results are often similar to criterion 7.
5. Kurtosis (difference spectra)	The kurtosis criterion is defined by the shape of the data-point distribution.	Although the kurtosis criterion is based on the distribution of data points, it is often difficult to use for the identification of deviations. Only in the case of an extremely high or negative value does it signify a deviation.
6. Correlation coefficient, linear regression (cumulative difference spectra)	The correlation coefficient determines the quality of a linear approximation. A value of $\sim$1 indicates a good approximation.	Typical XAS spectra differ too much from ideal linear behavior. Frequently the correlation coefficient does not clearly indicate deviant spectra.
7. Standard deviation, linear regression (cumulative difference spectra)	This criterion of linear regression calculates the standard deviation of each y value from the approximated linear function.	This criterion is often similar to the mean deviation but in addition accounts for the shape of the spectrum. In the investigation of several test systems this criterion gave the best results.
8. Slope, linear regression (cumulative difference spectra)	The slope is calculated by means of linear regression. For a flat line a slope of zero is expected.	While a large positive or negative slope is a good indicator for significant deviations, small values can also result from statistical cancellations (*e.g.* oscillations).

criterion 7 is sufficient for filtering out artefacts. Moreover, it is even easier to handle than the aforementioned criteria. Therefore we suggest the use of this criterion for quality control of XAS data.

2.7. Threshold determination

Statistical criteria can quantify variations and deviations in individual subsets. In any case, the aim of a quality control procedure is to reject only spectra with artefacts from the entire data set, while keeping the signal-to-noise ratio as high as possible. This requires the introduction of a threshold to estimate the overall quality of the data. Again criterion 7 is suitable. Its mean and maximum value for all subsets remaining in the pool can be used. Both values should decrease for each valid step in which a subset with deviations is eliminated. They have similar tendencies in this respect, but the changes of the mean value are typically more reliable indicators.

Upon removal of an individual contribution (or, if subsets are used, either one of the n rows or m columns of the $n \times m$ data matrix) both the mean and the maximum standard deviation will improve either because of the elimination of data exhibiting artefacts or due to statistical reasons. Even after all deviation-affected subsets are eliminated from the data pool there are still differences between the remaining subsets and their references. If such statistical variations are treated similar to significant deviations and the corresponding subsets are removed from the data pool, the procedure would continue until only a single subset is left. In order to avoid unnecessary restrictions of the data pool it is essential to define a threshold. This threshold has to differentiate between deviations which must be eliminated and variations without any significant effect on the data quality.

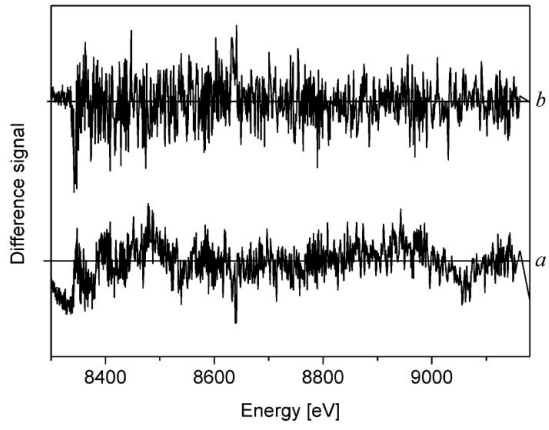

Figure 3
Example for a system [model compound Ni(btmg)SSiPh$_2$] where criterion 4 erroneously identifies the wrong spectrum while criterion 7 identifies the truly deviation-affected one. (*a*) The difference spectrum for channel 1 shows significant non-statistical deviations. Channel 1 is successfully indicated by criterion 7 as a potentially deviating channel. (*b*) The difference spectrum for channel 13 shows only statistical variations. Channel 13 is erroneously indicated by criterion 4 as a potentially deviating channel.

To verify the threshold, each elimination step is verified by comparison of the spectra defined by the resulting data sets. For the XAS data this verification is carried out by the comparison of the corresponding k^3-weighted EXAFS spectra as well as their Fourier transforms before and after each elimination step. The elimination of a subset with deviations should be indicated by differences between these spectra. In contrast, there should be no change of the overall shape of the EXAFS spectra or their corresponding Fourier transforms for the elimination of data which are only affected by statistical variations. Several different approaches using the mean and the maximum value of criterion 7 are tested for the definition of a suitable threshold.

If in a *simple approach* the mean or maximum value of criterion 7 is plotted for each elimination step, usually a decreasing asymptotic function with a certain limit determined by the horizontal asymptote is obtained (Fig. 4*a*). Although a threshold can be estimated when the curve approaches the asymptote, the exact position of the asymptote depends on the noise level of the data. Hence, no absolute value can be defined for the limit without extrapolation from the function.

The second strategy, here called the *static approach*, uses the ratio between the mean and the maximum value. If a subset with significant deviations is included in the data pool, the mean value of criterion 7 should be significantly lower than the maximum value – the value for a subset with potential deviations – resulting in a ratio of less than 1. When the maximum value approximates the mean value and the respective ratio approaches the limit of 1, a good overall data quality should be reached. The plot of the ratio for each step of the quality control procedure is a rising function (Fig. 4*b*). Unfortunately there is often no characteristic feature for this function on which the threshold can be based. In some cases the threshold was already reached for a ratio of less than 0.5. So with this approach there is the definite risk of eliminating too many spectra to reach a ratio close to the theoretical limit of 1. Therefore, the static approach should not be used for the determination of the threshold.

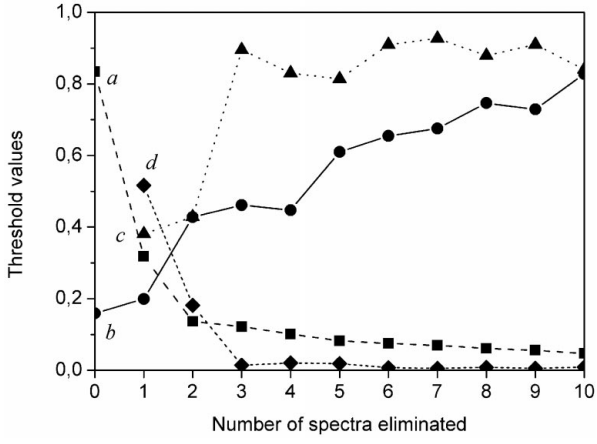

Figure 4
Data plots for different threshold approaches: (*a*) simple approach (mean value of criterion 7), (*b*) static approach (ratio of mean and maximum deviation of criterion 7), (*c*) dynamic approach (ratio of mean value of criterion 7 before and after the elimination step), (*d*) slope approach (slope between two data points of the simple approach).

In the *dynamic approach* the changes of the mean value of the selected criterion are used for the threshold determination. This is done by calculating the ratio between the mean value after and before the elimination step. For the elimination of a subset with deviations the value after the exclusion of the data should be significantly lower than before. The value of the corresponding ratio is therefore between 0 and 1. If the ratio for an iterative step is approximately 1, the elimination of the respective subset does not improve the quality of the data pool and is therefore not needed. Plotting the ratio for each elimination step results in an asymptotic function with a horizontal asymptote ($y = 1$) (Fig. 4*c*).

Although the plots for the simple and dynamic approach allow a relatively easy determination of a reasonable threshold, it is difficult to define a numerical argument that is required for automation.

For the data tested in the development of the quality control procedure a numerical argument based on the slope of the plot from the simple approach was found. If the absolute value of the slope between two subsequent data points was less than 0.1, the threshold was already reached and the corresponding elimination of a subset deemed unnecessary. The resulting plot for the *slope approach* (Fig. 4*d*) clearly indicates that for the example the threshold is reached after the elimination of the second channel.

Despite the fact that the *slope approach* worked well for the quality control of the investigated EXAFS data, it probably must be modified for other systems considering the amount of simultaneously examined spectra and the system-specific noise level of the data. In any case, with the guidelines discussed above, the threshold approaches can be easily adapted to other systems.

3. Example: quality control of XAS channel spectra

The quality control procedure has been tested for data obtained from different metalloproteins and model systems. Exemplarily quality control is presented for the screening of 13 channel spectra (Fig. 5) from Zn *K*-edge XAS measurements of a sample of *Glia cells missing* protein (GCM; Cohen *et al.*, 2002). Similar results were achieved for different samples of the NiFe hydrogenase from *Desulfovibrio vulgaris* MF and various nickel-containing model compounds (Lippold *et al.*, 2004).

3.1. Experimental

Measurements of the GCM sample were conducted at the EMBL EXAFS beamline D2 (HASYLAB, Hamburg) at 20 K. For data preparation and normalization the *EXPROG* software package (Nolting & Hermes, 1992) was used. The dead-time-corrected spectra were grouped into subsets according to the respective detector channels and for each channel the spectra of all suitable scans were averaged. The resulting 13 channel spectra were used for program-assisted quality

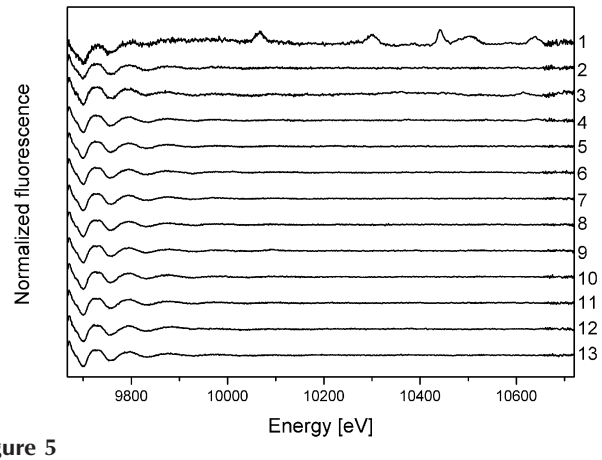

Figure 5
Comparison of the normalized raw spectra of GCM from each channel of the 13-element detector of the EMBL EXAFS beamline D2 (HASYLAB, Hamburg). To improve visibility only the fine structure above the absorption edge is shown.

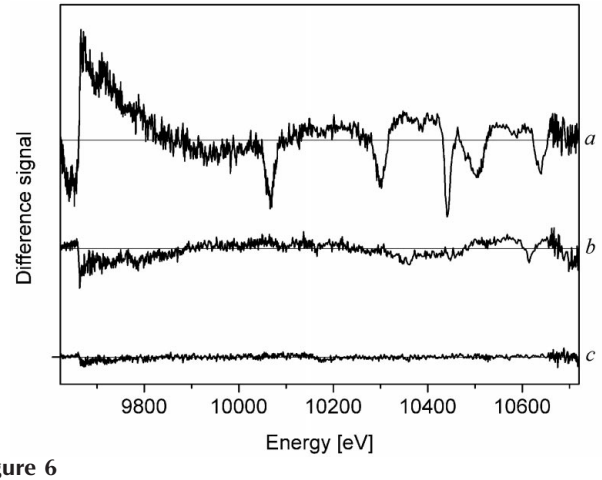

Figure 6
Comparison of the difference spectra for the quality control procedure of GCM. The spectra with the most significant deviations for each of the first three elimination steps are shown (*a*: channel 1; *b*: channel 3; *c*: channel 13).

control with the evaluation software *CHAOS* (channel analyzing and omitting system; Meyer, 2002; Lippold, 2003) based on spreadsheet calculations (StarCalc 6.0). Statistical criteria were automatically calculated and difference and cumulative difference spectra were plotted for comparison purposes. All aforementioned criteria were checked simultaneously and related to the plots of the difference spectra. The sequence in which individual channel spectra were identified as potential deviant spectra by the different criteria is presented in Table 3. The comparison with the visual inspection sequence provides a clear indication of the suitability of the criteria.

3.2. Iterative quality control

The quality control of XAS data from GCM is based on the standard deviation from linear regression (criterion 7) and verified by comparisons of the k^3-weighted EXAFS spectra and the corresponding Fourier transforms.

Table 3
Course of the iterative elimination of channel spectra for GCM based on different criteria.

Visible screening was only possible for two detector channels. Three of the remaining 11 elements exhibit similar small acceptable deviations. They are given in numerical order (2/12/13).

Criterion	Sequence of eliminated channel spectra
Visible screening	1, 3, 2/12/13
Largest positive deviation (1)	1, 3, 2, 4, 10, 8, 11, 12, 9, 7, 6, 13
Largest negative deviation (1)	1, 3, 2, 4, 6, 12, 13, 10, 11, 5, 7, 9
Summed positive deviations (2)	1, 3, 2, 12, 4, 11, 10, 8, 5, 9, 7, 6
Summed negative deviations (2)	1, 3, 2, 12, 13, 6, 4, 10, 9, 7, 5, 8
Summed deviations (3)	1, 2, 11, 12, 4, 8, 5, 7, 10, 9, 6, 13
Mean deviation (4)	1, 3, 2, 12, 13, 4, 10, 6, 9, 8, 11, 5
Kurtosis (5)	2, 6, 12, 13, 7, 5, 4, 9, 8, 1, 11, 3
Correlation coefficient, linear regression (6)	2, 3, 12, 4, 5, 8, 13, 11, 7, 9, 10, 6
Standard deviation, linear regression (7)	1, 3, 13, 12, 2, 11, 10, 6, 9, 4, 5, 8
Slope, linear regression (8)	1, 12, 2, 4, 5, 8, 11, 7, 9, 10, 13, 6

The first step of the iterative procedure for the contributions of the 13 channels is fairly straightforward. The deviations in the difference spectrum for channel 1 are clearly visible and most of the tested criteria indicate this channel as a possible deviant subset (Fig. 6*a*).

After the elimination of the data from channel 1 a significant improvement in the mean (62%) and maximum (70%) value of the standard deviation is achieved. The *slope approach* for the threshold determination is consistent with the elimination of this subset ($0.52 > 0.1$). The comparison of the k^3-weighted EXAFS spectra and their corresponding FT spectra shows significant changes when channel 1 is excluded from the data pool (Figs. 7*a* and 8*a*). This effect is most pronounced for the left shoulder of the main FT peak where an additional feature emerges. Therefore, channel 1 features significant deviations and must be eliminated from the data pool.

After the exclusion of channel 1 the next step of the iterative quality control procedure indicates channel 3 as a possible candidate for exclusion Again, this is illustrated by the difference spectrum (Fig. 6*b*). Besides its obvious bending there is a sharp feature at 10612 eV. If channel 3 is excluded from the data pool a significant but smaller improvement for the quality control criterion (57% for the mean value, 80% for the maximum value) results.

Comparison of the resulting EXAFS and Fourier transform spectra with and without channel 3 indicates small differences (Figs. 7*b* and 8*b*). Although the effects on the Fourier transform are less pronounced, there are significant differences for the k^3-weighted EXAFS spectrum in the region of $k > 15$. If the whole data range is used, channel 3 should therefore be excluded. This decision is supported by the threshold determination as the difference of the mean standard deviations amounts to 0.18 (>0.1).

The next iterative step identifies the contribution of channel 13 as a possible deviant spectrum. In contrast to the previous difference spectra its variations seem to be statistical (Fig. 6*c*). If channel 13 is removed from the data pool there is only a

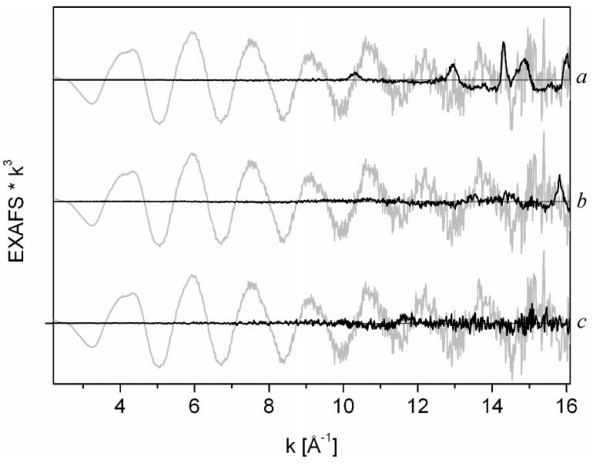

Figure 7
Comparison of the k^3-weighted EXAFS spectra for the quality control procedure of channel spectra from XAS measurements of GCM. EXAFS spectra including the potentially erroneous channels (grey line) and the difference between the spectra including and excluding the potentially erroneous channels (black line) are plotted for the elimination steps of the first three channels (*a*: channel 1; *b*: channel 3; *c*: channel 13).

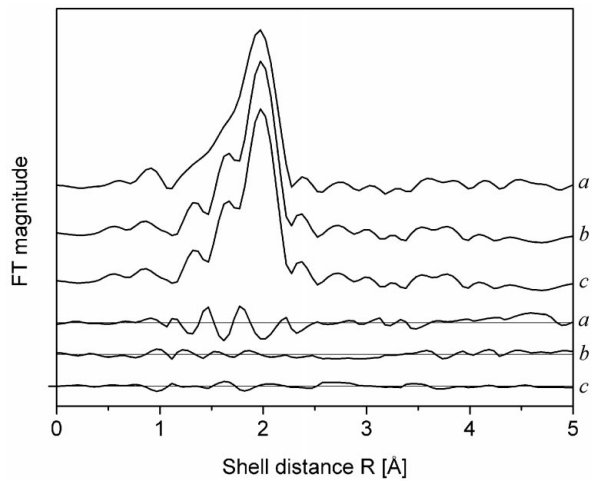

Figure 8
CComparison of the Fourier transform spectra for the quality control procedure of channel spectra from XAS measurements of GCM. Fourier transform spectra including the potentially erroneous channel (upper part) and the difference between the spectra including and excluding the potentially erroneous channel (lower part) for the elimination steps of the first three channels and the corresponding difference plot are shown (*a*: channel 1; *b*: channel 3; *c*: channel 13).

slight improvement of the standard deviation (10% for the mean value, 17% for the maximum value). In this case the differences between the k^3-weighted EXAFS spectra and their respective Fourier transforms appear to consist purely of statistical variations (Figs. 7*c* and 8*c*). The effects on the Fourier transform are negligible for the whole spectrum.

The calculated value for the *slope approach* amounts to 0.01 and is significantly lower than the limit of 0.1. Thus this data elimination step is unnecessary as there is no improvement to the quality of the data pool.

Consequently, the channel spectrum 13 appears to be valid and the threshold is reached after elimination of two channel spectra. As the data from channel 13 feature the strongest variations of the remaining spectra it can be assumed that contributions from all other channels can be included as well in the final data set. Hence, only the spectra of channels 1 and 3 have to be excluded.

Although the threshold is apparently reached, the procedure is reiterated for verification purposes until only three channel spectra remain. With each step no significant effect on the EXAFS and FT spectra is observed (despite the increase of the noise level). This verifies the quality control procedure as well as the threshold. The plots of the criteria for threshold determination defined by the simple, static, dynamic and slope approach are shown in Fig. 4.

For the exemplary quality control of XAS channel spectra of GCM both criterion 7 and the slope approach to threshold determination are successful in identifying deviating spectra. Similar results were achieved for data from other biological samples and model compounds.

4. Conclusion

In this contribution several criteria and iterative procedures for the quality control of XAS spectra were tested and applied to BioXAS data. By the introduction of difference spectra between the data subsets and the remaining data pool and by transformation into cumulative difference spectra, deviations can be identified using a simple criterion from linear regression. For online feedback to the data acquisition this system can in principle be supplemented by other criteria to gain information on the type of deviation. The identification of deviant spectra can be successfully verified using k^3-weighted EXAFS spectra and their corresponding Fourier transforms.

To avoid unnecessary reduction of the data pool a threshold for the elimination of spectra is introduced. While the approach to quality control presented here was successfully applied to the EXAFS data used for the development of the quality control procedure, modifications might be necessary to account for other specific systems. As different approaches are discussed, adaptation for other data sets should be easy.

The resulting quality control procedure *CHAOS* using spreadsheet algorithms was successfully applied to spectra from the EMBL EXAFS beamline D2 (DESY, Hamburg) as a test system.

Because of the simplicity of the different steps of the procedure, the resulting quality control system can easily be adapted and used for the quality control of XAS data from different sources. Thus the quality control system presented in this publication allows for the first time the automated quality control of BioXAS data and XAS data from other dilute samples.

Financial support from the Bundesministerium für Bildung und Forschung (BMBF) (grant No. 05 KS1PGA/5) is gratefully acknowledged.

References

Abramowitz, M. & Stegun, I. A. (1972). Editors. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, 9th printing. New York: Dover.

Ascone, I., Fourme, R. & Hasnain, S. S. (2003). *J. Synchrotron Rad.* **10**, 1–3.

Ascone, I., Meyer-Klaucke, W. & Murphy, L. (2003). *J. Synchrotron Rad.* **10**, 16–22.

Cohen, S. X., Moulin, M., Schilling, O., Meyer-Klaucke, W., Schreiber, J., Wegner, M. & Müller, C. W. (2002). *FEBS Lett.* **528**, 95–100.

Edwards, A. L. (1976). *An Introduction to Linear Regression and Correlation.* San Francisco: W. H. Freeman.

Lippold, B. (2003). *CHAOS: Channel Analyzing and Omitting System.* Universität Duisburg-Essen, Germany.

Lippold, B., Fichtner, C., Lubitz, W., Meyer-Klaucke, W. & Henkel, G. (2004). In preparation.

Meyer, T. (2002). PhD thesis, Gerhard-Mercator-Universität Duisburg, Germany. Lilienthal: Vlg. Simmering.

Nolting, H.-F. & Hermes, C. (1992). *EXPROG: EMBL EXAFS Data Analysis and Evaluation Package for PC/AT*. EMBL Hamburg, Germany.

Ranieri-Raggi, M., Raggi, A., Martini, D., Benvenuti, M. & Magani, S. (2003). *J. Synchrotron. Rad.* **10**, 69–70.