Journal of
Synchrotron
Radiation

ISSN 0909-0495

Received 15 September 2003
Accepted 5 November 2004

New methods for EXAFS analysis in structural genomics

Grant Bunker,* Nicholas Dimakis and Gocha Khelashvili

Illinois Institute of Technology, Physics, BCPS Department, Chicago, IL 60616, USA.
E-mail: bunker@iit.edu

Data analysis is one of the remaining bottlenecks in high-throughput EXAFS for structural genomics. Here some recent developments in methodology are described that offer the potential for rapid and automated XAS analysis of metalloproteins.

Keywords: XAS; EXAFS; data analysis; structural genomics.

1. Introduction

The Human Genome Project has provided a wealth of genetic information whose structural and functional implications have only begun to be tapped. A large amount of effort is presently being invested to express, crystallize, measure diffraction patterns and determine the crystal structures of expressed proteins.

Even having accomplished this, determining the functional role of a protein from a crystal structure is generally not a simple or even well defined task. In many cases proteins may require insertion of metal cofactors or other post-translational modifications that occur in cells, but may or may not happen *in vitro*. Proteins may fold differently depending on the metal cofactors with which they are presented. Evidently there is a need to structurally characterize metalloprotein structures before, during and after crystallization.

X-ray absorption fine-structure (XAFS) spectroscopy has strengths and weaknesses relative to X-ray crystallography and higher-dimensional NMR. Although it is a relatively near-sighted technique, probing atoms only several angstroms from the selected metal atoms, if the metal is in a functionally important site within the protein it provides a unique perspective, and can do so under non-crystalline conditions.

In recent years substantial improvements have been made in third-generation synchrotron radiation sources, beamline optics, controls and detectors. Good quality EXAFS data on millimolar concentration samples can be acquired with scan times of seconds and total integration times of minutes. Analyzers are now available (Zhang *et al.*, 1999; Karanfil *et al.*, 2000) that effectively eliminate problems due to detector saturation. Cumulatively these advances offer the potential for high-throughput acquisition of XAFS data for structural genomics.

Despite these improvements to experimental methods, however, a significant rate-limiting step still exists: analysis of the EXAFS data. Historically this has been carried out in a labor-intensive manner by data reduction and subsequent modeling of the atomic distribution. There have been significant improvements in theoretical methods (Ankudinov *et al.*, 1998; Filipponi & Di Cicco, 2000), and with improvements in computational power there now appears to be a good potential for high-throughput methods of data analysis of biological data.

2. Data reduction and analysis

Biological EXAFS data analysis can be divided into three stages: data reduction, first-shell data analysis and multi-shell modeling. Data reduction normally consists of the following steps: applying instrumental (*e.g.* detector dead-time) corrections if needed; scan averaging to improve signal-to-noise ratio; selection of energy zero; normalization to unit edge step; interpolation to k-space; and background subtraction to obtain the EXAFS signal. These steps can be carried out in an automated manner (Newville *et al.*, 1993).

First-shell data analysis is usually accomplished by restricting the range of interest in r-space by Fourier methods. Modeling the first-shell distribution of atoms may be simple or complex depending on the number of different atomic species present in the first shell, the variation in distances, and more subtle issues such as the relative scattering phases of the atoms, the extent of the usable data range, weighting *etc*. The traditional approach of modeling the distribution as sums of Gaussian or slightly non-Gaussian (*e.g.* Poisson distribution) subshells is often successful, but this non-linear modeling process can be poorly conditioned or even numerically unstable in complex cases. Because the fitting is non-linear, *i.e.* the fitting function depends non-linearly on the fit parameters, traditional algorithms are prone to get caught in multiple false minima rather than the global best fit.

3. Direct methods to determine the nearest-neighbor distribution

There now exist (Khelashvili & Bunker, 1999; Khelashvili, 2001) direct methods for determining the distribution of atoms in the first shell in an automated way. These do not require the

analyst to assume a particular functional form of the distribution.

To a good approximation the strong single-scattering signals can be separated from the weak first-shell multiple-scattering signals based on their path length. Reconstruction of the interatomic distance distribution amounts to solving a set of linear equations, which by virtue of their linearity have a single global minimum. However, the solution is numerically unstable unless other sufficient constraints or *a priori* information are included. This inverse problem [reconstruction of the radial distribution function (RDF) from the experimental EXAFS data] can be 'regularized' (made stable) by inclusion of smoothness constraints, and normalization constraints if appropriate. Early work (Babanov *et al.*, 1981) introduced regularization methods to EXAFS analysis. Recently this approach has been extended (Khelashvili & Bunker, 2004; Babanov, 2004) to automate the choice of regularization parameters and also to generate error bars for the generated RDFs.

By way of illustration we consider a notoriously (Clark-Baldwin *et al.*, 1998) ill-behaved experimental situation that occurs in XAFS when nitrogen or oxygen (*e.g.* His or Asp/Glu) ligands are simultaneously present with sulfur (*e.g.* cysteine) ligands in some unknown configuration. The difficulty lies in the fact that the phases of the N/O and S signals are approximately 180° out of phase over the data range, so that, in the fitting process, S atoms appear roughly as 'anti-nitrogen/oxygen'. This results in destructive interference and a high degree of parameter correlation between the coordination numbers for N and S if both are allowed to float, as well as exacerbating the usual correlation between the Debye–Waller factors (DWF) and coordination numbers. This is one of many cases in which a good understanding of the non-linear fitting problem is needed. Unacceptable results will be obtained if traditional methods are automatically (or naively) used.

Fig. 1 shows a calculated spectrum for a situation in which there are four N atoms at 2.00 Å and two S atoms at 2.20 Å, with random noise added to simulate experimental noise in the data. The single-scattering amplitudes and phases for nitrogen and sulfur were generated using *FEFF*8.0 (Ankudinov *et al.*, 1998). The synthetic noise was generated so as to increase in proportion k^3, with a maximum value of 0.5 at k = 16. The projected Tikhonov Landweber–Friedman (PTLF) regularization method was then used to invert the data into two RDFs, one for nitrogen and one for sulfur. This automated direct method does not assume any particular functional form of the RDFs and it also handles well distributions that include non-Gaussian disorder. Fig. 2 shows the reconstructed RDFs with error bars that are estimated by the method. The integrals of the RDFs yield the correct coordination numbers of four N and two S and distances, even without imposing any other constraints on the RDFs. Including *a priori* information, if available (*e.g.* total coordination number of six), reduces the error bars further. The primary limitation at this point is that the approach is limited to single-scattering analysis.

In the context of structural genomics, the PTLF XAFS method affords the possibility of automatically determining the distributions of mixed nearest-neighbor atoms, such as oxygen/nitrogen *versus* sulfur ligands in a relatively high throughput mode. This nearest-neighbor information can be used to supplement and cross-check results from X-ray crystallography and NMR, and to check for correct metal insertion and folding of the protein before crystallization.

Inclusion of information from atoms beyond the nearest coordination shell can be very illuminating for structure determination by XAFS. For example, it is well known that multiple scattering from second- and third-shell atoms of the imidazole ring of coordinated histidine residues contributes characteristic features to the EXAFS spectra. More than two decades ago this was used to determine the presence and geometry of coordinated histidine residues in the iron site of

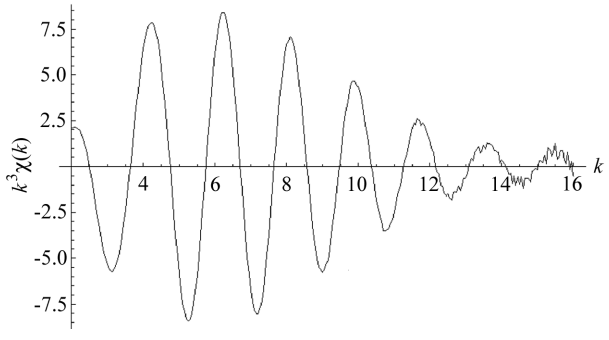

Figure 1
EXAFS spectrum for four N atoms at 2.00 Å, σ^2 = 0.003 Å^2, and two S atoms at 2.20 Å, σ^2 = 0.005 Å^2, calculated using *FEFF*8.0 and with simulated noise.

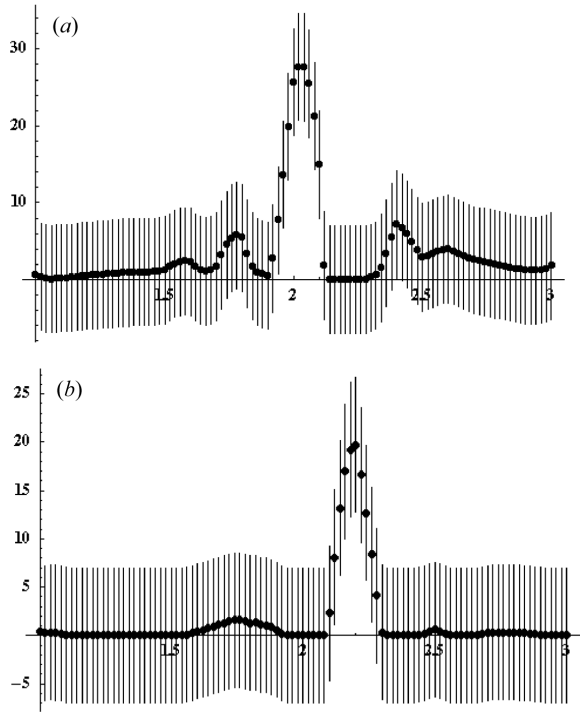

Figure 2
Radial distribution functions for (*a*) nitrogen and (*b*) sulfur as reconstructed by PTLF regularization.

photosynthetic reaction centers (Bunker *et al.*, 1982), without benefit of any prior crystallographic analysis.

There is considerable information in the more distant coordination shells, but also much greater complexity. Multiple scattering (MS) of the photoelectron must be explicitly considered in most cases; unlike single scattering, multiple scattering depends on the three-dimensional geometry of the metal site. Fortunately in recent years there has been excellent progress in developing programs to theoretically calculate MS XAFS spectra, taking as input a known or hypothetical three-dimensional structure. These theoretical calculations can be used to fit experimental data and determine structures, provided the effects of vibrations on the spectra are also accurately accounted for.

Vibrational effects are often described in terms of the EXAFS DWFs, which are similar to but distinct from the analogous quantities in X-ray diffraction. EXAFS DWFs depend on the mean square variation of the path length for all important multiple-scattering paths. These can number in the hundreds, so that it may be impossible to fit them without exceeding the limited information content of the data. If the force constants for relevant bonds can be accurately determined by alternative means, or are sufficiently few in number that they can be treated as floating variables, it is possible to directly fit the EXAFS data.

In practice, the parameters have often been specified in an *ad hoc* manner by making plausible estimates of the DWFs and verifying that they give good results on known test cases. This is not a very satisfactory procedure, however, and better alternatives are desirable. Dimakis and Bunker (Dimakis *et al.*, 1999; Dimakis & Bunker, 2001) have explored this issue and have recently found that calculating the force constants by density functional theory (DFT) can prove multiple-scattering DWFs that are reliable and accurate. The force 'constants' are rather strongly dependent on bond length, being dependent on chemistry, and tabulated force constants were insufficient for the purpose. The DWFs of all important paths have been mapped out for histidine, cysteine and carboxylate residues as a function of interatomic bond length, angle (as necessary) and sample temperature. Structures such as Zn tetraimidazole, as shown in Fig. 3, are accurately calculated from first principles with no free parameters.

One essential point is that in this approach the laborious DFT calculations need to be performed and parameterized only once, and the results can then be reconstructed rapidly and easily for use in fitting experimental data. As of the third quarter of 2003, tables are presently available for Zn sites, Cu sites are presently being parameterized, and the other first-row transition-metal sites are planned, contingent on funding.

Another key point is that it is possible (Dimakis & Bunker, 2002) to model the spectra on a group-by-group basis, *i.e.* apart from scattering through the central metal atom the inter-group multiple scattering is negligible to a first approximation and only intra-group scattering is included. This is important because it is presently not practical to calculate the vibrational dynamics of the whole metal active site by DFT for all hypothetical conformations.

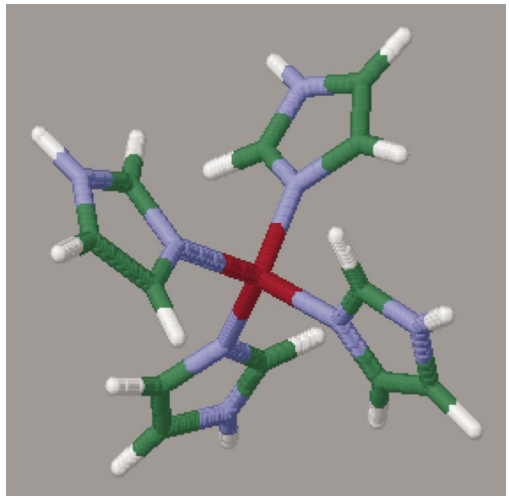

Figure 3
Structure of Zn tetraimidazole, test case for *ab initio* Debye–Waller calculations (Dimakis *et al.*, 1999; Dimakis & Bunker, 2001).

4. Inverse problem for multishell modeling

The foregoing section describes one viable approach to the DWF problem in the multiple-scattering fitting of biological XAFS data. There may be other viable approaches, but it is clear that the technical infrastructure can be put in place for rapid multi-shell calculation of accurate EXAFS spectra, including vibrational effects. This 'forward problem' (calculating the spectrum accurately if the structure is known) is only part of the problem, however. Data analysis requires solution of the inverse problem: determining the structure from the data.

The inverse problem is actually not well posed. In general there may be multiple structures that are consistent with the data, within the experimental (and computational) uncertainties. The analyst's job is to identify and describe the entire range of structural models that are consistent with the data, within the errors. Well chosen *a priori* information and constraints can be vital for excluding unphysical solutions and for stabilizing the inversion. A multiple-scattering generalization regularization procedure for single-scattering analysis has not yet been developed, so indirect methods, *e.g.* non-linear fitting, must presently be used.

Fortunately there are advances in this area as well. Simple yet powerful algorithms are available for global minimization that can exploit the inexpensive computational power provided by parallel computers. The simplest approach involves job farming of conventional minimization algorithms with different randomly selected starting values. Different starting values may result in minimization into either the same or different local minima, and these 'attractors' within the parameter space can be mapped out and characterized according to their goodness of fit. This generally involves many function evaluations. However, with a number of modern inexpensive computers, rapid solution of unknown structures is possible because the approach is inherently parallel and asynchronous and requires only moderate-bandwidth interconnects between the computers. Other parallel

algorithms such as parallel genetic algorithms (Storn & Price, 1995; Rabinowitz, 1995) and hybrid genetic/Levenberg–Marquardt approaches are also promising for this purpose, and have been successfully applied to fitting XAFS. The implementation of these on larger clusters is a subject of future research. Constraints based on *a priori* information can be imposed through penalty functions, and the parameter errors estimated in a self-consistent manner by Bayesian methods (Krappe & Rossner, 1999).

Because of the generality of this approach, fitting can also be used in concert with crystallographic data in simultaneous refinement (Hasnain & Strange, 2003). For this approach to be valid, the XAFS spectra and the diffraction data must be measured on the same crystalline forms, and the effects of X-ray linear dichroism must be explicitly addressed, and preferably used to advantage.

5. Conclusion

The primary bottlenecks for both rapid and robust experimental data acquisition and analysis have been eliminated in principle. A dedicated effort to implement these approaches in an integrated system for XAFS structural genomics could be a viable complement to macromolecular crystallography and NMR.

Funding for much of this work was provided by BioCAT through National Institutes of Health (NIH) grant RR08630. Additional support was provided from the State of Illinois Higher Education Cooperation Act.

References

Ankudinov, A. L., Ravel, B., Rehr, J. J. & Conradson, S. D. (1998). *Phys. Rev. B*, **58**, 7565–7576.
Babanov, Yu. A. (2004). To be published.
Babanov, Yu. A., Vasin, V. V., Ageev, A. L. & Ershov, N. V. (1981). *Phys. Status Solidi B*, **105**, 747.
Bunker, G., Stern, E. A., Blankenship, R. & Parson, W. W. (1982). *Biophys. J.* **37**, 539–551.
Clark-Baldwin, K., Tierney, D. L., Govindaswamy, N., Gruff, E. S., Kim, C., Berg, J., Koch, S. A. & Penner-Hahn, J. E. (1998). *J. Am. Chem. Soc.* **120**, 8401–8409.
Dimakis, N., Al-Akhras, M.-A. & Bunker, G. (1999). *J. Synchrotron Rad.* **6**, 266–267.
Dimakis, N. & Bunker, G. (2001). *J. Synchrotron Rad.* **8**, 297–299.
Dimakis, K. & Bunker, G. (2002). *Phys. Rev. B*, **65**, 201103-1–201103-4.
Filipponi, A. & Di Cicco, A. (2000). *Task Q. 4*, **4**, 575–669.
Hasnain, S. S. & Strange, R. (2003). *J. Synchrotron Rad.* **10**, 9–15.
Karanfil, C., Zhong, Z., Chapman, L. D., Fischetti, R., Bunker, G. B., Segre, C. U. & Bunker, B. A. (2000). *Am. Inst. Phys. Conf. Proc.* **521**, 178–182.
Khelashvili, G. (2001). PhD dissertation, Illinois Institute of Technology, USA.
Khelashvili, G. & Bunker, G. (1999). *J. Synchrotron Rad.* **6**, 271–273.
Khelashvili, G. & Bunker, G. (2004). To be published.
Krappe, H. J. & Rossner, H. H. (1999). *J. Synchrotron Rad.* **6**, 302–303.
Newville, M., Livins, P., Yacoby, Y., Stern, E. A. & Rehr, J. J. (1993). *Phys. Rev. B*, **47**, 14126–14131.
Rabinowitz, F. M. (1995). *ACM Trans. Math. Softw.* **21**, 194–213.
Storn, R. & Price, K. (1995). *Differential Evolution – A Simple and Efficient Adaptive Scheme for Global Optimization over Continuous Spaces.* Technical Report TR-95–012. Berkeley: International Computer Science Institute.
Zhang, K., Rosenbaum, G. & Bunker, G. (1999). *J. Synchrotron Rad.* **6**, 220–221.