Journal of
Synchrotron Radiation

ISSN 0909-0495

Received 14 September 2003
Accepted 5 November 2004

Bottlenecks and roadblocks in high-throughput XAS for structural genomics

Robert A. Scott,[a]* Jacob E. Shokes,[a] Nathaniel J. Cosper,[a] Francis E. Jenney[b] and Michael W. W. Adams[b]

[a]Department of Chemistry, University of Georgia, Athens, GA 30602, USA, and [b]Department of Biochemistry and Molecular Biology, Department of Chemistry, University of Georgia, Athens, GA 30602, USA. E-mail: scott@chem.uga.edu

Structural and functional characterization of the entire protein complement (the proteome) of an organism can provide an infrastructure upon which questions about biological pathways and systems biology can be framed. The technology necessary to perform this proteome-level structural and functional characterization is under development in numerous structural genomics and functional genomics initiatives. Given the ubiquity of metal active sites in a proteome, it seems appropriate to ask whether comprehensive local structural characterization of metal sites within a proteome (metalloproteomics) is either a valid or obtainable goal. With a proteome-wide knowledge of the active-site structures of all metalloproteins, one could start to ask how metal insertion, cluster assembly and metalloprotein expression are affected by growth conditions or developmental status *etc.* High-throughput X-ray absorption spectroscopy (HTXAS) is being developed as a technology for investigating the metalloproteome. In creating a pipeline from genome to metalloproteome, several bottlenecks to high-throughput determination of metal-site structures must be overcome. For example, automation of arraying small samples for XAS examination must be invented, automation of rapid data collection of multiple low-volume low-concentration samples must be developed, automation of data reduction and analysis must be perfected. Discussed here are the promises and the pitfalls of HTXAS development, including the results of initial feasibility experiments.

Keywords: metalloenzymes; metalloproteomics; proteomics; X-ray absorption fine structure.

1. Introduction

The promise of the genomic revolution will only be realised as gene sequence information is interpreted in terms of gene product (*e.g.* protein) structure and function. Determination of structure–function relationships within whole proteomes requires development of high-throughput technologies, an area of significant effort and investment over the last few years. Structural genomics initiatives are hoped to deliver enough coverage of sequence-structure space that prediction of gene product structure from gene sequence becomes possible. (This is equivalent to solving the 'protein folding problem'.) The next step is prediction of function from structure with the ultimate paradigm being immediate functional prediction upon sequencing a new gene (an outcome of functional genomics).

In the near term, however, the rate and efficiency of generation of protein structural information is a major bottleneck to entering this post-genomic era (NIGMS, 2001). This is true for the traditional tools of structural biology, X-ray crystallography and solution structure determination by multi-dimensional nuclear magnetic resonance (NMR) techniques. In this article, we evaluate the state of the art in the use of a non-traditional structural biology tool, X-ray absorption spectroscopy (XAS), applied at the proteome level. XAS determination of local structural details of metal (and some non-metal) sites provides a useful complement to the more traditional structural biology approaches and can also yield unique electronic structural information that probes active site catalysis. We argue that high-throughput (HT) XAS is a valid technology to develop (in parallel with HT crystallography *etc.*) to target the 'metalloproteome', the set of all metal-site structures within components of the proteome. We also discuss the technological developments that will be required to develop HTXAS, as well as some initial feasibility studies.

2. Metalloproteomics

We use this term herein to refer to the biological information to be obtained from HTXAS determination of metal-site structures within the proteome. The metalloproteome is distinct from the metallome, defined as the distribution and level of metals throughout a cell or organism (Williams, 2001).

What sorts of questions can be answered using HTXAS and how will this advance the biology of metalloproteomics? Fig. 1 summarizes two potential workflows for characterizing the metalloproteome from a given genome. Given the need to examine homogeneous metal-site structures with XAS, HTXAS will certainly take advantage of separation technologies, including those developed for proteomics and structural genomics. The top sequence in Fig. 1 describes the separation at the gene level. As in most current structural genomics efforts, individual gene products are generated by HT cloning, expression and purification. XAS characterization is then carried out on individual metalloprotein samples. The bottom sequence in Fig. 1 delays the separation step to the proteome stage. After expression of the entire proteome (from microorganisms grown under specific conditions, or the protein complement of cells of a given organ or tissue at a specific stage of development, for example), separation of proteins generates the individual metalloprotein samples to be characterized by XAS.

By either route, the collection of individual proteins, in addition to being identified by standard proteomics techniques (*e.g.* mass spectrometry), is subjected to XAS analysis to investigate first metal content, then metal-site structure. Fig. 1 suggests a staged analysis in which initial interrogation of the samples by X-ray fluorescence (XRF) provides proteomic metal distribution. Measurement of XANES spectra provides information about oxidation state and 'fingerprint' speciation of metal sites. Finally, a full EXAFS analysis yields the desired metal-site structures that are the raw data of the metalloproteome.

This technology applies to either workflow sequence of Fig. 1. The application of this technology to address biologically interesting questions depends on which sequence is used. The top workflow generally depends on cloning individual genes with purification tags, heterologous expression and affinity purification, allowing the automation necessary for high throughput. Biological information about relative expression levels, post-translational processing and modification *etc.* is lost in this workflow. Still, determination of metal-site structures formed by a given protein could be used to build a database of metal-binding sequence motifs, a 'motif taxonomy'. Such a database would be an important component of a future sequence-to-structure predictor. Given the importance of metals in many biological processes, functional genomics will require a consideration of metal binding sites.

More information about metals in biology is available through the bottom workflow of Fig. 1. The metal sites found in proteins expressed *in vivo* are expected to retain biological relevance. Distribution of metals and relative expression levels of metalloproteins as a function of environment (*e.g.* microbial growth conditions) could be determined readily. But the promise of metalloproteomics goes beyond the usual expression profiling; this is accessible by standard proteomics techniques, in which the polypeptide chain of a given metalloprotein is quantified and related to environmental influences. Formation of biologically relevant metal sites often requires a set of post-translational events, including metal transport, trafficking, metallocenter assembly, cluster exchange *etc.*, any of which could be influenced by environmental changes. These metal-based post-translational events are invisible to standard proteomics analysis and are therefore the sole purview of metalloproteomics. Of special interest are effects of metals in the environment: what metals bind to which proteins in what coordination upon essential metal depletion, toxic metal levels *etc*? It is access to this level of information that makes development of HTXAS desirable.

3. Feasibility studies

Each step of each workflow in Fig. 1 must be adapted to provide high throughput for HTXAS to succeed. Before discussing how this pipeline might be designed, we discuss XAS studies at 'low throughput' to test the feasibility of performing XAS on the type of sample we expect to use. In these feasibility studies we take advantage of access to a large collection of expressed genes from the Protein Production Module of the Southeast Collaboratory for Structural Geno-

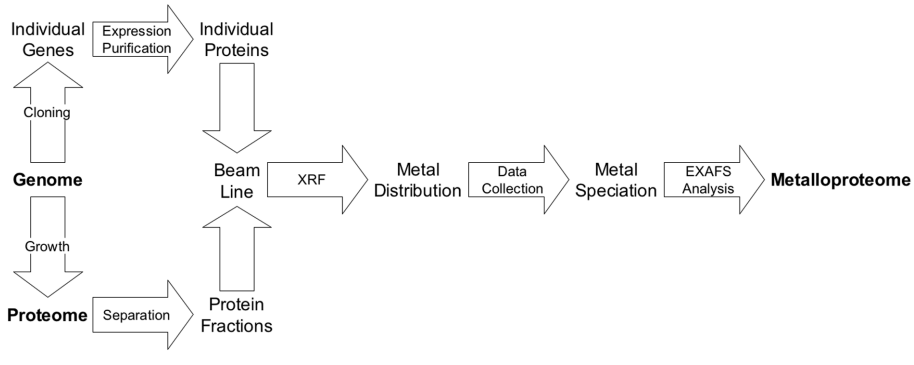

Figure 1
Schematic flowchart for the two alternate workflows for high-throughput X-ray absorption spectroscopy (HTXAS). The top workflow involves separation of genes by cloning, followed by HT expression and purification prior to organizing multiple samples for further analysis. The bottom workflow involves expression of the proteome prior to separation and organizing the multiple samples. Sample throughput can be accomplished either by spatial arraying or automated temporal sequencing. Synchrotron radiation analysis consists of mapping the metal distribution using X-ray fluorescence (XRF), using XANES for metal speciation and using EXAFS for metal-site structural analysis of the metalloproteome. Each arrow indicated on the flowchart represents a potential bottleneck for the overall high-throughput pipeline.

mics (SECSG) (Adams *et al.*, 2003). A significant fraction of the ~2200 open reading frames in the *Pyrococcus furiosus* (*Pf*) genome have been individually cloned, tagged, expressed and purified by high-throughput techniques at levels necessary for crystallography and NMR analysis. By comparison, XAS analysis requires relatively small amounts; our initial studies utilized ~3 µL samples of 0.2–1.0 m*M* protein.

As an initial test of high-throughput methodology, we designed a 25-well sample holder with 1.5 mm-diameter holes in a 5 × 5 arrangement on a 1" wide polycarbonate holder that fits into the liquid-helium-flow XAS cryostat in use at the Stanford Synchrotron Radiation Laboratory (SSRL). Up to 25 *Pf* ORF product samples were loaded (3 µL per well) and the sample holder frozen in liquid nitrogen before insertion into the cryostat. This is only one potential method for achieving high sample throughput. One could also imagine a robot-automated single-sample positioning device analogous to those being used for high-throughput X-ray diffraction data collection currently (Lesley *et al.*, 2002; Adams *et al.*, 2003).

With the beam apertured to 1 mm × 1 mm, the cryostat and sample holder were rastered first to align the wells, then the metal distribution was determined by monitoring the K_α X-ray emission from multiple elements at once using a multi-channel energy-discriminating solid-state fluorescence detector. For example, with an excitation (monochromator) energy set above the Zn *K* edge (10200 eV), the distributions of Co, Ni, Cu and Zn were mapped simultaneously during one raster scan. These elemental maps were used to target protein samples for further speciation (by XANES) and structural analysis (by EXAFS). Fig. 2 shows the quality of such data obtained on BL 9-3 at SSRL on SPEAR 2.

4. Bottlenecks in the pipeline

The workflows summarized in Fig. 1 can be considered a pipeline for structural information from the genome to the metalloproteome. Each arrow in Fig. 1 represents a potential bottleneck in this pipeline. Successful development of HTXAS requires careful attention to removing these bottlenecks. Current structural genomics efforts have already streamlined the cloning, expression and purification of gene products (the top workflow of Fig. 1) to provide samples for XAS investigation (Lesley *et al.*, 2002; Adams *et al.*, 2003). The critically important bottom workflow of Fig. 1 has received much less attention, at least at the production level needed for either HTXAS or more traditional structural genomics tools (crystallography, NMR). There is no question that new developments in separation technologies would contribute to the success of this version of HTXAS. For example, standard proteomics application of denaturing 2*d* electrophoresis is useless, since most metals are lost upon denaturation. Improvements in native 2*d* gel separations may prove useful, as would alternative medium-scale separation methods.

We will concentrate here on XAS-related bottlenecks. Improving throughput for the initial determination of metal distribution requires mostly technical improvements. For example, if samples are to be arrayed in holders (as described above), sample loading can be carried out robotically. Raster scanning of the wells can be accelerated and identification of metals in wells can probably be automated (by appropriate analysis of multi-channel analyzer traces), allowing data collection at multiple edges on multiple samples to proceed without manual intervention. Alternatively, if frozen solution samples of individual gene products can be handled like protein crystals, then sample-handling robots that are being developed now for structural genomics (Lesley *et al.*, 2002; Adams *et al.*, 2003) can be used to sequentially load samples, identify metals and collect XAS data quickly and without manual intervention. The software used to control data collection would of course have to be modified from that used in current structural genomics programs.

The rest of the HTXAS workflow would also require attention. Data collection needs to happen rapidly and will probably utilize recent developments in 'quick-EXAFS'

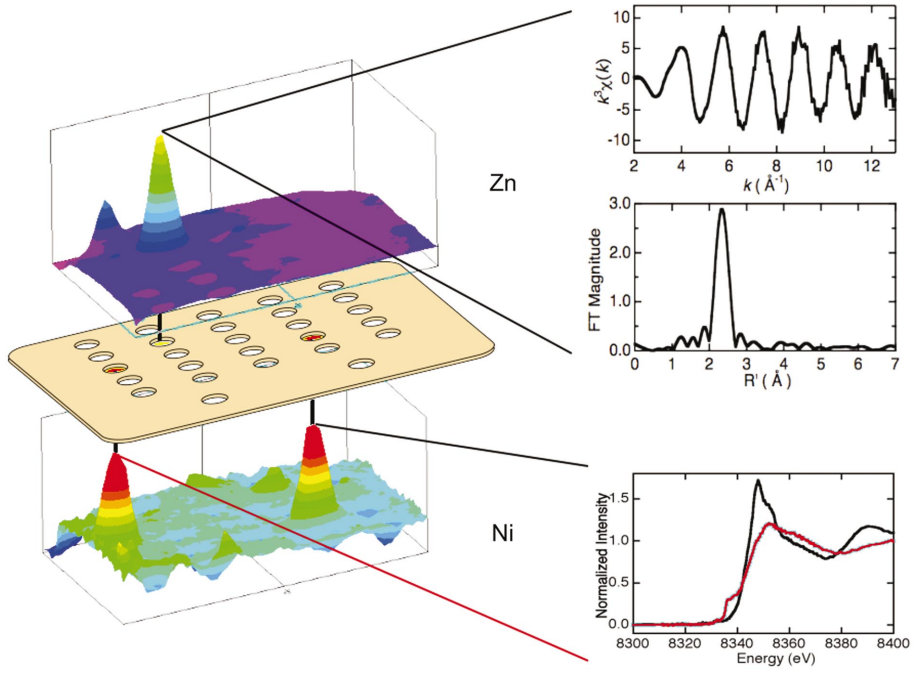

Figure 2
Examples of the data collected to test the feasibility of the top workflow of Fig. 1. Individual gene product samples from the *Pyrococcus furiosus* genome were loaded into a 5 × 5 spatial array of 3 µL samples (middle left). A single raster scan was used to monitor both Ni (lower left) and Zn (upper left) K_α emission, identifying the presence of these metals in the indicated wells. Ni XANES speciation (lower right) clearly differentiates the types of Ni binding sites and Zn EXAFS structural analysis (upper right) shows the sensitivity available for determining metal-site structures. The element maps and XAS data were collected on beamline 9-3, SSRL (SPEAR 2), using a 1 mm × 1 mm focused beam and 30-element intrinsic germanium detector.

technology (Richwin *et al.*, 2002). Finally, data reduction and data analysis require significant development effort. For a 'first look' at the metal site, some form of automated data reduction (calibration, background subtraction, EXAFS extraction) will be useful, as will at least a crude automated first-shell curve-fitting analysis. With a relatively limited scatterer set of biologically relevant ligands, a relatively rapid yet reliable initial prediction of possible first-shell ligands should be available. More detailed EXAFS analysis on most of the metalloproteome will be desirable and tools need to be developed to make this less time-consuming. Spending days or weeks refining curve-fitting optimization for a single sample (as happens sometimes now) is not possible when this is considered at the proteome level (hundreds or thousands of data sets).

5. Conclusion

Metalloproteomics, as a discovery science, will go beyond standard proteomics by providing information about how an organism adjusts its post-translational machinery to manufacture metal-containing active sites depending on environmental signals and stresses. Development of the technology (high-throughput X-ray absorption spectroscopy) to investigate the metalloproteome will require a coordinated multidisciplinary effort in biochemistry, physics, engineering and informatics. This investment will not only provide return in our improved understanding of the biological pathways of metal trafficking, but in improved technologies for all XAS efforts, whether they are high-throughput or not.

Support for X-ray absorption spectroscopy research in the RAS laboratory is supported by the National Institutes of Health (GM 42025). The Southeast Collaboratory for Structural Genomics is supported by grants from the National Institutes of Health (GM 62407), the Georgia Research Alliance and the University of Georgia. XAS data were collected at the Stanford Synchrotron Radiation Laboratory, a national user facility, operated by Stanford University on behalf of the US Department of Energy, Office of Basic Energy Sciences. The SSRL Structural Molecular Biology Program is supported by the Department of Energy, Office of Biological and Environmental Research, and by the National Institutes of Health, National Center for Research Resources, Biomedical Technology Program.

References

Adams, M. W., Dailey, H. A. DeLucas, L. J., Luo, M., Prestegard, J. H., Rose, J. P. & Wang, B.-C. (2003). *Acc. Chem. Res.* **36**, 191–198.

Lesley, S. A., Kuhn, P., Godzik, A., Deacon, A. M., Mathews, I., Kreusch, A., Spraggon, G., Klock, H. E., McMullan, D., Shin, T., Vincent, J., Robb, A., Brinen, L. S., Miller, M. D., McPhillips, T. M., Miller, M. A., Scheibe, D., Canaves, J. M., Guda, C., Jaroszewski, L., Selby, T. L., Elsliger, M. A., Wooley, J., Taylor, S. S., Hodgson, K. O., Wilson, I. A., Schultz, P. G. & Stevens, R. C. (2002). *Proc. Natl. Acad. Sci. USA*, **99**, 11664–11669.

NIGMS (2001). *The Structures of Life*, NIH Publication 01–2778. National Institute of General Medical Sciences, Bethesda, MD, USA.

Richwin, M., Zaeper, R., Lützenkirchen-Hecht, D. & Frahm, R. (2002). *Rev. Sci. Instrum.* **73**, 1668–1670.

Williams, R. J. P. (2001). *Coord. Chem. Rev.* **216**, 583–595.