Characterization of metalloproteins by high-throughput X-ray absorption spectroscopy

Wuxian Shi,[1,7] Marco Punta,[2] Jen Bohon,[1] J. Michael Sauder,[3] Rhijuta D'Mello,[1] Mike Sullivan,[1] John Toomey,[1] Don Abel,[1] Marco Lippi,[4] Andrea Passerini,[5] Paolo Frasconi,[4] Stephen K. Burley,[3] Burkhard Rost,[2,6] and Mark R. Chance[1]

[1]New York SGX Research Center for Structural Genomics (NYSGXRC), Case Western Reserve University, Center for Proteomics and Bioinformatics, Case Center for Synchrotron Biosciences, Upton, New York 11973, USA; [2]TU Munich, Informatik, Bioinformatik, Institute for Advanced Studies, 85748 Garching, Germany; [3]New York SGX Research Center for Structural Genomics (NYSGXRC), Eli Lilly and Company, Lilly Biotechnology Center, San Diego, California 92121, USA; [4]Dipartimento di Sistemi e Informatica, Università degli Studi di Firenze, 50139 Firenze, Italy; [5]Dipartimento di Ingegneria e Scienza dell'Informazione, Università degli Studi di Trento, 38123 Povo, Italy; [6]New York Consortium on Membrane Protein Structure, New York Structural Biology Center, New York, New York 10027, USA

High-throughput X-ray absorption spectroscopy was used to measure transition metal content based on quantitative detection of X-ray fluorescence signals for 3879 purified proteins from several hundred different protein families generated by the New York SGX Research Center for Structural Genomics. Approximately 9% of the proteins analyzed showed the presence of transition metal atoms (Zn, Cu, Ni, Co, Fe, or Mn) in stoichiometric amounts. The method is highly automated and highly reliable based on comparison of the results to crystal structure data derived from the same protein set. To leverage the experimental metalloprotein annotations, we used a sequence-based de novo prediction method, MetalDetector, to identify Cys and His residues that bind to transition metals for the redundancy reduced subset of 2411 sequences sharing <70% sequence identity and having at least one His or Cys. As the HT-XAS identifies metal type and protein binding, while the bioinformatics analysis identifies metal- binding residues, the results were combined to identify putative metal-binding sites in the proteins and their associated families. We explored the combination of this data with homology models to generate detailed structure models of metal-binding sites for representative proteins. Finally, we used extended X-ray absorption fine structure data from two of the purified Zn metalloproteins to validate predicted metalloprotein binding site structures. This combination of experimental and bioinformatics approaches provides comprehensive active site analysis on the genome scale for metalloproteins as a class, revealing new insights into metalloprotein structure and function.

[Supplemental material is available for this article.]

Some 13% of all proteins of known three-dimensional structure possess a bound metal. Examination of the Protein Data Bank (PDB; http://www.rcsb.org) reveals that Mg and Zn are the most abundant, while Ca, Mn, Fe, and Ni are also frequently observed. Metalloproteins represent one of the most diverse classes of proteins, with the intrinsic metal atoms providing catalytic, regulatory, or structural roles critical to protein function (Degtyarenko 2000). For example, Zn, the most abundant metal in cells, plays a vital role in the function of more than 300 enzyme classes, in stabilizing the DNA double helix and in control of gene expression (Andreini et al. 2006).

Metalloproteomics is a relatively new field addressing genome-scale identification and functional analysis of metal-associated proteins (Szpunar 2005; Bertini and Cavallaro 2008; Shi and Chance 2008). Several major experimental approaches have been successfully employed in metalloproteomics, including both forward and reverse technologies. Reverse technologies follow classical biochemical approaches, wherein samples are fractionated by sequential chromatographic or 2-D gel separations, and "fractions" are subjected to metal and protein identification techniques, such as atomic absorption and mass spectrometry (LC-MS and/or ICP-MS) to identify the metal and protein associated with specific fractions (Bettmer 2005; Szpunar 2005; Kulkarni et al. 2006; Manley et al. 2009; Bartel et al. 2010; Cvetkovic et al. 2010). An advantage of this approach includes isolation of protein from a native cell or tissue environment; potential disadvantages are potential loss of native metal and difficulties in generating sufficient quantities of the protein of interest for high signal-to-noise analysis. The forward approach described herein is also derived from classic methods and involves cloning and expression of the genes of interest followed by analysis of metal content and function. An advantage of this approach is the ability to optimize the expression and amount of the protein of interest; disadvantages include possible nonnative metal incorporation or loss of native metal due to expression protocols, although strategies for metal exchange to optimize spectroscopic analysis are well-known and extremely valuable (Chance et al. 2004; Scott et al. 2005; Shi et al. 2005). In addition, computational approaches can complement these experimental methods and explore wide ranges of sequence space for their potential connections to metal binding and related enzyme functions (Andreini et al. 2009). Both experimental approaches suffer from potential misannotation of a metalloprotein in two distinct ways. First, when native metals are lost in purification,

[7]**Corresponding author.**
E-mail wushi@bnl.gov.

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.115097.110.

a bound metal may not be associated with the protein (false negative). Second, nonnative metals may bind in the place of native metals, which may mislead the investigator with respect to the native and functionally active metalloprotein species (false positive). Such challenges have been well-understood in the metalloprotein field for many years, as confirmation of the role of a metal in a protein's native function requires careful experimental work. By combining a range of computational approaches with either forward or reverse experimental annotations, genome scale metalloprotein annotation is under active exploration by many research groups (Shi et al. 2005; Bertini and Cavallaro 2008; Cvetkovic et al. 2010).

In this paper, we describe a high-throughput forward approach taking advantage of purified proteins arising from the U.S. Protein Structure Initiative structural genomics effort. As such, we have access to thousands of expression constructs and purified proteins and are able to directly measure the metal content of each protein utilizing high-throughput X-ray absorption spectroscopy (HT-XAS). HT-XAS employs the physical principle of X-ray excitation of core electrons of metal atoms with detection of resulting X-ray fluorescence as in X-ray absorption spectroscopy (Shi et al. 2005). However, metal identification and quantification are based on the energy and intensities of the X-ray fluorescence signal emitted by intrinsic metals bound to the proteins, respectively. HT-XAS in metalloproteomics so far has been closely associated with structural genomics projects (Scott et al. 2005; Shi et al. 2005).

One aim of the U.S. structural genomics program has been to experimentally determine at least one protein structure from each protein family ≤30% sequence identical to any protein with a known 3-D structure, thus enabling structural analysis of the entire family through the use of comparative modeling. Targets are selected based on this overall strategy, and protein structures are determined by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy (Shi and Chance 2006; Manjasetty et al. 2007). The success rate from cloning through purified protein to the final experimental 3-D structures by the large-scale structural genomics centers in the NIH-funded Protein Structure Initiative (PSI) is ~10%–14% (Bonanno et al. 2005; Graslund et al. 2008; Sauder et al. 2008); however, the set of purified proteins as a whole represents a useful resource for biochemical and biophysical characterization of the proteins to better understand their functions.

Previously, we reported a HT-XAS study of a limited ensemble of proteins (a total of 654) from PSI Phase 1 (PSI-1) via quantitative analysis of six metals (Mn, Fe, Co, Cu, Ni, and Zn) (Chance et al. 2004; Shi et al. 2005). Protein samples were purified using the high-throughput protein production pipeline of the New York Structural Genomics Research Consortium (NYSGXRC) (Bonanno et al. 2005; Sauder et al. 2008). Over 10% of the total samples showed the presence of transition metal atoms in stoichiometric amounts. The method was shown to be over 90% accurate in predicting the presence or absence of a transition metal based on 50 crystal structures from the sample set. Bioinformatics-based functional annotation was carried out for the metalloproteins identified by HT-XAS. In many cases, the metal binding information provided a distinct new annotation for proteins of unknown function and improved annotation for proteins with poorly understood function.

We report herein extension of the method to nearly 4000 purified proteins generated by the NYSGXRC during PSI Phase 2 (PSI-2). Automation of the method was improved to handle thousands of samples. Within the PSI structure determination pipeline, the results were used to assist in structure solution by X-ray crystallography, primarily in the map interpretation step where the metal identities are sometimes ambiguous. Proteins annotated as metal-binding were analyzed by computational methods including PSI-BLAST (Altschul et al. 1997) for detection of similar sequences that may be used for annotation transfer via homology and by MetalDetector (http://metaldetector.dsi.unifi.it) (Passerini et al. 2006; Lippi et al. 2008), a de novo method that predicts His and Cys residues that bind to transition metals using information derived from protein sequence. Structural models available from homology modeling of (often distant) templates deposited in MODBASE (Pieper et al. 2004) or built utilizing SWISS-MODEL (Kiefer et al. 2009) were also used to identify putative metal-binding sites, and these models were evaluated in the context of experimental and bioinformatics results to provide a comprehensive understanding of the metalloprotein structure of the targets. Overall, our results show that the information from both experimental and bioinformatics approaches can be productively combined to improve our understanding of metal-binding sites in metalloproteins.

Results and Discussion

Identification of metalloproteins

We analyzed 3879 proteins from NYSGXRC during the period from September 2006 to August 2009. Three hundred forty-three protein samples (8.8% of the proteins in total) contained at least one of the detectable metals (Table 1), including Mn, Fe, Co, Ni, Cu, and Zn. We established a threshold of 0.3 for the metal-to-protein molar ratio, at or above which a target can be assigned as a "valid" metalloprotein, while the cutoff in our previous study was assigned as 0.5 (Shi et al. 2005). We reduced the cutoff for these data for the following reasons. First, we improved the signal-to-noise with the current experimental setup, thereby providing more confidence when measuring relatively low fluorescence signals. Second, a significant percentage of the protein samples included related proteins (i.e., distinct proteins derived from different expression constructs of the same target, different protein preparations from the same construct, or a homologous protein from a different organism). HT-XAS results from such related proteins can be used corroboratively. For example, target 9276b, a protein from the amidohydrolase superfamily, was analyzed three times with samples expressed both in *Escherichia coli* (9276b1BCt7p1) and insect cells (9276b4KWg2h1, 9276b5KWg2h2). All three measurements indicated the presence of Zn with a Zn/protein ratio of 0.2, 0.3, and 0.3, respectively. With such reproducible results, the protein can be safely assigned as capable of Zn binding. In addition, target 9276d (a homolog from a different species) was analyzed twice by HT-XAS, with Zn detected in one measurement and Mn/Fe detected in another. Based on these results, we speculate that 9276d, which is similar to 9276b, is also a Zn metalloprotein. For the X-ray crystallographer, this information might be important for obtaining diffraction quality crystals (by adding metal ions to the crystallization buffers), for phasing using the intrinsic metal as an

Table 1. Statistics of HT-XAS of NYSGXRC targets from 2006–2009

	Proteins	Metalloproteins	Zn	Cu	Ni	Co	Fe	Mn
Total	3879	345	94	9	33	3	133	140
Percentage		8.9	2.4	0.2	0.9	0.1	3.4	3.6

anomalous scatterer, and for resolving metal ambiguities in the electron density map interpretation step. In particular, for this protein, small amounts of Zn can be added to protein buffers prior to crystallization to increase Zn occupancy and maximize the protein homogeneity for optimal diffraction.

From 343 metalloproteins, Zn was detected 94 times, whereas Fe and Mn were each detected ~140 times (Table 1). Mn and Fe were frequently detected at the same time in the same sample. Examination of the PDB indicates that Zn and Mg are the most abundant metals, followed by Ca, Mn, Ni, and Fe (Shi et al. 2005), although we are unable to detect Ca or Mg using our approach. The unusually high frequencies of Mn and Fe detected in the large set of proteins are due to the target selection strategy of NYSGXRC in PSI-2 that has some bias toward the detailed study of community- nominated targets. During the second phase of PSI (July 1, 2005–June 30, 2010), the four Large-Scale Production Centers were required to commit 70% of their effort on centrally chosen targets designed to expand coverage of "sequence-structure space," 15% effort on targets nominated by the broader research community, and 15% effort on specific biomedical themes chosen independently by each center. Many of the NYSGXRC community-nominated targets are from two large metal-dependent protein superfamilies, the amidohydrolases and enolases (Pieper et al. 2009). Although the substrates vary, the reactions catalyzed by enzymes in the enolase superfamily share a common core chemical step of an abstraction of a proton from a carbon adjacent to a carboxylic acid and an essential divalent metal ion (Babbitt et al. 1996; Yew et al. 2006). Enolases typically prefer Mg^{2+} as the cofactor, but Mg^{2+} can be substituted by other divalent metals, such as Mn^{2+} or Fe^{2+}. Characterization of Mg^{2+} is not currently implemented in our HT-XAS procedure, as the low energy of its fluorescence signal makes its detection challenging; however, many targets from the enolase superfamily were found to contain partially occupied Mn^{2+} and/or Fe^{2+}, in some cases likely binding to the proteins opportunistically during protein expression or purification, while in other cases representing the native binding.

Overall agreement between HT-XAS and MetalDetector predictions

To benchmark our experimental annotation against computational predictions, we analyzed the protein sequences represented by the purified proteins using the algorithm MetalDetector. The program uses machine-learning techniques to predict transition metal binding for His (two states: metal bound or free) and Cys (three states: metal bound, free, or disulfide bonded) residues in protein sequences. Binding to other residues such as Asp or Glu cannot be easily predicted by machine-learning methods due to the paucity of available examples and the fact that their background binding probability (e.g., the ratio of metal-bound-Glu/all-Glu) is low (Passerini et al. 2006; Shu et al. 2008). As HT-XAS detects binding of a specific metal to the protein, while MetalDetector predicts binding of a transition metal atom (with the type not specified) to specific residues within a sequence, the computational results add significant value to the experimental metal annotation and can be compared in detail for the degree of overlap between predictions under varying computational parameters and experiments.

Although MetalDetector was not developed (i.e., optimized) to predict metal binding at the protein level, in addition to reporting on specific prediction examples throughout the paper, we wanted to examine the overall agreement between MetalDetector and HT-XAS in the identification of likely metalloproteins. In order to use MetalDetector to predict metal binding at the protein level rather than at the His/Cys residue level, we developed a simple prediction scheme. We predicted a target as a metalloprotein if the metal-bound state output score (mbss), i.e., the score that relates to the probability of a residue to be in a metal-bound state, of at least N among the His and Cys residues exceeding a given threshold (T_M; see Methods). For example, if the threshold T_M = 0.4 and N = 2, we would predict a target to be a metalloprotein if at least two among its His and Cys residues featured an mbss > 0.4. Note that we could predict a residue to be metal-bound in this scheme even if its free state score is higher than the mbss score; mbss need only be higher than the threshold, T_M. In general, since several His and Cys residues often combine to form transition metal-binding sites, we expected that the presence in a protein of multiple residues with mbss above a given T_M threshold might result in a more reliable set of predictions.

In Figure 1A, we report relative-precision–relative-recall curves (see Methods for definitions) for N = 1 and N = 5, calculated by moving the threshold on the mbss value from 1 to 0 while analyzing a 70% redundancy reduced set of HT-XAS- analyzed proteins. At 10% relative-recall, relative-precision goes from 42 ± 4% for N = 1 to

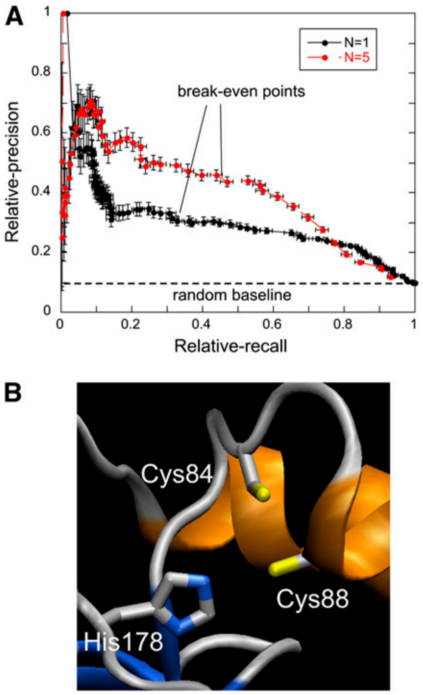

Figure 1. His and Cys metal binding predictions using the program MetalDetector (Passerini et al. 2006; Lippi et al. 2008). (*A*) Relative-precision–relative-recall curves for MetalDetector predictions. N is the number of His and Cys residues that have an mbss value exceeding a given T_M threshold. N = 1 (black) and N = 5 (red). The random baseline is calculated as the fraction of all metalloproteins identified by HT-XAS divided by all proteins. Averages and standard errors are calculated using bootstrapping without resampling (Efron and Tibshirani 1993). (*B*) Putative metal-binding site in target 11211f formed by Cys84 (metal = 0.47; free = 0.2; disulf = 0.33), Cys88 (metal = 0.56; free = 0.09; disulf = 0.36), and His178 (metal = 0.41; free = 0.59). Model obtained with SWISS-MODEL in alignment mode (Kiefer et al. 2009). As a template, we used the Cu^+-binding NMR structure of human SCO2 (PDB ID: 2RLI) (Banci et al. 2007), which shares 26% sequence identity with target 11211f (alignment length is 161, using three iterations of PSI-BLAST). VMD (Humphrey et al. 1996) was used for molecular graphics.

66 ± 5% for $N = 5$. Because of the way relative-precision and relative-recall are defined, the break-even point between them is also the point at which MetalDetector predicts about the same number of metal-binding proteins as HT-XAS. This occurs at different values of T_M depending on N. Relative-precision and -recall at the break-even point goes from 32% for $N = 1$ ($T_M = 0.41$) to ~45% for $N = 5$ ($T_M = 0.14$). In other words, when HT-XAS and MetalDetector identify the same number of metalloproteins, they agree on one-third to close to one-half of the cases, depending on N. This data can be compared to a random predictor baseline of 10% obtained by dividing the overall number of HT-XAS-annotated metalloproteins by all proteins analyzed. MetalDetector relative-precision approaches the random baseline only at high recall. MetalDetector also outperforms predictions obtained by calculating the random baseline in a different, less naive way (see Supplemental Fig. S1).

It must be emphasized that MetalDetector metalloprotein predictions that are not identified or confirmed by HT-XAS are expected as, in many cases, metal occupancy is low, and metal atoms are not detected by HT-XAS. For example, proteins 11211f and 11213j share the Pfam SCO1/SenC domain (PF02630) involved in biogenesis of respiratory and photosynthetic systems. Neither of the two proteins was identified by HT-XAS to bind metal. In both sequences, MetalDetector assigned fairly high mbss values to one His and two Cys residues. In 11211f (GenBank AAU26510), these residues are Cys84 (mbss = 0.47), Cys88 (mbss = 0.56), and His178 (mbss = 0.41); the highest mbss for any other His or Cys in the protein is equal to 0.07. Note that while Cys84 mbss is higher both with respect to the free-state score (0.2) and to the disulfide-bonded state score (0.33) (this is trivially true also for Cys88), in the case of His178 the free-state score (0.59) is actually higher. Predictions for the corresponding 11213j residues are similar but slightly less significant. Human SCO2 is a mitochondrial membrane-bound protein involved in copper supply for the assembly of cytochrome c oxidase and shares 26% sequence identity over 161 residues with 11211f. The NMR structure of the human protein has a Cu^+ ion bound (Banci et al. 2007). In the alignment with 11211f, the His and the two Cys found in the metal-binding site of Human SCO2 correspond to Cys84, Cys88, and His178 (Fig. 1B). 11211f and 11213j likely constitute a case in which MetalDetector recovered a metalloprotein missed by HT-XAS. Note that a homolog of SCO2 (24% and 22% sequence identity with 11211f and 11213j, respectively), Sco1 from *Bacillus subtilis*, was present in the MetalDetector training set, increasing the prospects of identifying a similar protein. Despite the low overall sequence similarity, this likely helped in predicting 11211f metal-binding residues.

Zn metalloproteins

A total of 94 protein samples from NYSGXRC showed the presence of Zn (ratio ≥ 0.30). Approximately half of these proteins have a measured Zn/protein ratio of ≥0.50, indicating that they most likely have one or more Zn-binding sites. Forty Zn metalloproteins, which represent different NYSGXRC targets, were chosen for detailed bioinformatics-based functional annotation. A BLAST search to identify related protein families and PDB entries was conducted on these 40 targets, and the results are shown in Supplemental Table S1 and Table 2. This analysis indicates close relationships between the selected targets and Zn-containing families and is highly supportive of the experimental annotation of metal binding by HT-XAS.

Of the 40 proteins identified in Supplemental Table S1, 35 were found in the 70% redundancy reduced data set defined above for comparing HT-XAS results and MetalDetector predictions. For $N = 1$ and at $T_M = 0.41$ (i.e., at the $N = 1$ MetalDetector break-even point of the relative-precision–relative-recall curve on the whole reduced set), relative-recall on these 35 proteins is 54 ± 5%, i.e., much higher than the one for the whole redundancy reduced set at the break-even point value (32 ± 2%). This result is not surprising, as Zn is better represented in the available PDB training data and, as a result, MetalDetector predicts Zn binding better than for most other transition metals (Passerini et al. 2006).

For illustration, we examine two examples of Zn metalloproteins identified by HT-XAS in detail (Table 2). HT-XAS results indicate that NYSGXRC target 10382a (GenBank NP_390104 residues 3-749) contains a zinc/protein ratio of 0.6. The protein is annotated as a putative ATP-dependent helicase, and a BLAST search showed that the protein belongs to a family of helicases with a metal-binding cysteine cluster. MetalDetector predicts 4 cysteine residues in a cluster [Cys718 (mbss = 0.92), Cys720 (mbss = 0.93), Cys724 (mbss = 0.98), and Cys727 (mbss = 0.96)], each residue having a high probability of binding to a transition metal. Among the remaining His and Cys residues, the one with the highest mbss is His721 (mbss = 0.22). In this case, there was no match between 10382a and proteins in the MetalDetector training set at a BLAST E-value < 10^{-3}. On the other hand, a BLAST search against the PDB found a few remotely related helicases (e.g., PDB IDs: 2ZJ2, 1OYY, 2V1X, with maximal sequence identity of 23%). Alignments to these helicases, however, cover only about two-thirds of 10382a and do not contain its predicted C-terminal metal-binding cysteine cluster (Bernstein et al. 2003; Oyama et al. 2009; Pike et al. 2009).

In general, more than half of the Zn proteins in Supplemental Table S1 have a related crystal structure in the PDB (27/40). For the four targets from the amidohydrolase superfamily, 9328a, 9256a, 9247a, and 9236e, crystal structures have been determined and deposited in the PDB (PDB IDs: 3GUW, 2IMR, 2QEE, and 3HPA, respectively; Table 2), and all include a bound Zn (3GUW includes two Zn ions). For these proteins, MetalDetector does not return strong metal-binding predictions, although for target 9256a it predicts a number of His residues as part of the Zn-binding site with a relatively high mbss: His101, mbss = 0.36; His240, mbss = 0.47; and the nearby His303, mbss = 0.34.

Some targets, instead, have only a remotely related structure or a structure for a single domain. For example, the three sequence-related targets—12087a (putative uncharacterized protein; Zn, Ni/protein ratios = 0.9, 0.5, respectively; Table 2), 12087b (SIR2 family transcriptional regulator; Zn/protein ratio=0.4), and 12087c (putative uncharacterized protein; Zn/protein ratio=0.5)—exhibit low but significant sequence similarity to a number of PDB structures (30% sequence identity, using BLAST), including Sir2 protein PDB ID: 1ICI (Min et al. 2001). SIR2 proteins are deacetylases that depend on nicotine dinucleotide (NAD) (Finn et al. 2008). In the crystal structure, a Zn ion that seems to play a mainly structural role is bound to two cysteine clusters, conserved in the three target sequences (except that 12087c lacks one of the conserved cysteines in the first motif). 1ICI was part of the MetalDetector training set, and in all three cases, MetalDetector predicts that the two cysteine clusters participate in metal binding also in the NYSGXRC targets (Fig. 2; residue numbers and mbss found in the caption are for the best-predicted protein, 12087b, GenBank ABB30640 residues 2-278): Cys140 (mbss = 0.84); Cys144 (mbss = 0.55); Cys170 (mbss = 0.98); and Cys173 (mbss = 0.93). Note that in the homology model, Cys144 faces away from the putative binding site due to a one-residue insertion in 12087b with respect to the template

Table 2. Selected Zn metalloproteins identified from NYSGXRC PSI-2 proteins and annotation of closely related genes

ID	Metal	NMA	Length	Protein annotation	Clusters of orthologs	BLAST-PDB	Related PDB
12087a	Zn,Ni	0.9, 0.5	285	AC:AAM99938.1; hypothetical protein; OS: *Streptococcus agalactiae*	Ev = 2 × 10^{-131}; NAD-dependent protein deacetylases,SIR2 family	Transcriptional regulatory protein, SIR2 family; from *Archaeoglobus fulgidus*	PDB ID: 1ICI; 16% identity; metal ion = Zn
9328a	Zn	1	249	AC:gi|11499354, hypothetical protein AF1765; OS: *Archaeoglobus fulgidus*	Ev = 1 × 10^{-136}; metal-dependent hydrolase, TatD-related deoxyribonuclease	Ev = 1 × 10^{-136}; Tatd-like protein (Af1765); from *Archaeoglobus fulgidus*	PDB ID:3GUW; identities = 97%; metal ion = Zn
9256a	Zn	1.6	418	AC:gi|15805850; putative hydrolase; OS: *Deinococcus radiodurans*	Ev = 0.0; metal-dependent hydrolase, amidohydrolase	Ev = 0.0, amidohydrolase Dr_0824; from *Deinococcus radiodurans*	PDB ID: 2IMR; identities = 100%; metal ion = Zn
9247a	Zn	0.8	426	AC:gi|10173106; putative amidohydrolase; OS: *Bacillus halodurans*	Ev = 0.0; amidohydrolase	Ev = 0.0; putative amidohydrolase BH0493; from *Bacillus halodurans*	PDB ID:2QEE; identities = 100%; metal ion = Zn, Mg
9236e	Zn	0.6	478	AC:gi|44264246; amidohydrolase; OS:unknown	Ev = 0.0; hydroxydechloroatrazine ethylaminohydrolase	Ev = 0.0; amidohydrolase; from an environmental sample from the Sargasso Sea	PDB ID: 3H4U; identities = 100%; metal ion = Zn
10382a	Zn	0.6	747	AC:P50830; putative ATP-dependent helicase; OS: *Bacillus subtilis*	Ev = 0.0; helicase family protein with metal-binding cysteine cluster	Ev = 5 × 10^{-22}; archaeal DNA helicase	PDb ID: 2ZJ2; identities = 25% (436 residues); metal ion = no

(NMA) Number of metal atoms per protein molecule; (AC) accession number; (OS) organism/species; (Ev) E-value from BLAST search.

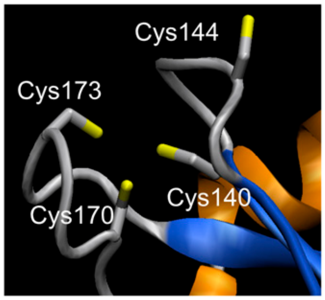

Figure 2. Putative Zn^{+2}-binding site in target 12087b formed by Cys140 (metal = 0.84; free = 0.02; disulf = 0.14), Cys144 (metal = 0.55; free = 0.15; disulf = 0.3), Cys170 (metal = 0.98; free = 0; disulf = 0.02) and Cys173 (metal = 0.93; free = 0.01; disulf = 0.06). Model obtained with SWISS-MODEL in alignment mode (Kiefer et al. 2009). As a template, we used the Zn^{+2}-binding crystal structure of the Sir2 protein from *Archaeoglobus fulgidus* (PDB ID: 1ICI) (Min et al. 2001), which shares 22% sequence identity with target 12087b (alignment length is 272, using three iterations of PSI-BLAST). VMD (Humphrey et al. 1996) was used for molecular graphics.

(likely responsible for the lower mbss of this Cys with respect to the others). In the actual 12087b structure, we would probably expect the protein backbone to rearrange in such a way that Cys144 is found in a favorable position for Zn binding. In this case (12087b), MetalDetector additionally predicts a few His residues with a relatively high mbss but only one His with mbss > 0.5 (His132; mbss = 0.65). In this example, HT-XAS data, homologs found in the PDB and sequence-based predictions all support the hypothesis that 12087b is a metal- and possibly Zn-binding protein.

Comprehensive approaches

Our long-term goal for the project is to establish a metalloprotein database associated with the PSI Knowledgebase (http://kb.psi-structuralgenomics.org/) that will include information such as metal content, metal-binding site residue predictions, and a 3-D structure of the metal-binding site based on homology models. Toward this end, we have shown how the experimental annotations of metal assignment are leveraged by the computational predictions of binding residues by MetalDetector. We now provide two examples illustrating how these data can be collectively incorporated in the context of homology models to provide structure parameters for predicted metal-binding sites.

The first example is a NYSGXRC PSI-1 target (not included in the set analyzed in the previous paragraphs), T812/892d (Fig. 3A GenBank gi 16079612). HT-XAS screening showed that the protein contained Zn with a 0.5 Zn/protein ratio. The protein is a putative cytidine deaminase, based on sequence similarity. MetalDetector identified three residues, His70, Cys98, and Cys101, with mbss 0.42, 0.80, and 0.73, respectively, suggesting that this protein most likely binds metal (the highest mbss for any other His or Cys in this protein is 0.18 for His178; no match for T812/892d with proteins in the MetalDetector training set at BLAST E-value < 10^{-3}). A homology model, built from a template with moderate sequence similarity (PDB ID: 1KGD, 35% identity, no metal), was retrieved from MODBASE. These three putative metal-binding residues are clustered together to form a potential Zn-binding site. In addition, another His residue, His95, is located within 5 Å, most likely contributing to Zn coordination. In this case, although the homology model did not provide a metal annotation, the union of bioinformatics and experimental data can provide support for inclusion of a metal atom in a specific structural site, thereby strengthening confidence in the resulting homology model.

The second example is a NYSGXRC PSI-2 target, 10060e (GenBank gi 46578500), a protein of unknown function from *Desulfovibrio vulgaris* (Fig. 3B). Ni was identified through HT-XAS screening, providing a metalloprotein annotation. We retrieved from MODBASE a homology model of the protein, built from a distantly related structure determined by NYSGXRC (PDB ID: 2IJQ, 20% identity, no metal). In our homology model, four residues, His32, His35, Asp36, and Tyr63, were found to be <6 Å away from each other. In this case, MetalDetector assigns low mbss scores to both His32 and His35 (0.02 and 0.21, respectively); at the same time, it predicts an unrelated His, His65, with a slightly higher mbss (0.27). The amino acid types identified through the homology model are frequently observed in Zn-binding sites. Thus, we speculate that the protein is a Zn metalloprotein and that the Zn atom is likely replaced by Ni during protein purification due to a His-tag strategy involving a Ni^{2+} chelating column.

The approaches described here—namely experimental identification of metal content, prediction of residues involved in metal binding, and 3-D modeling of the metal-binding site—combine to provide a powerful and complementary approach to studying the structure and function of metalloproteins.

Testing of structure models using extended X-ray absorption fine structure (EXAFS)

In order to determine whether proposed metal-binding sites predicted through the comprehensive approach could be validated, EXAFS experiments were conducted on two Zn-binding proteins, with target identifiers 9550a (gi 15025762) and 9453d (gi 8777583) (MetalDetector predicts one His with mbss = 0.35 in 9550a and two His, both with mbss = 0.48, in 9453d, although not all of them

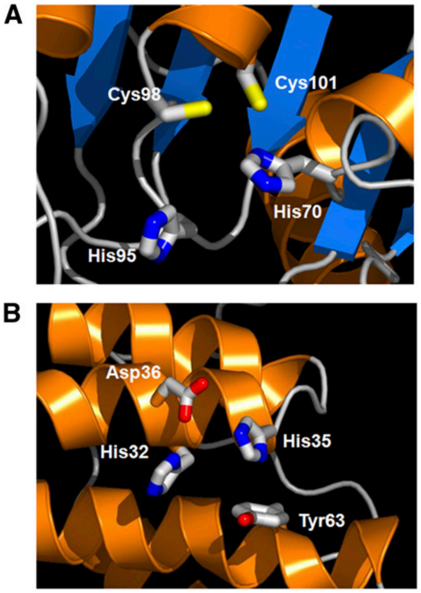

Figure 3. Examples of complementary use of HT-XAS and homology modeling. (*A*) Zn-binding site for target NYSGXRC-892d (aka T812). MetalDetector (Passerini et al. 2006) identified three residues, His70, Cys98, and Cys101, with mbss of 0.42, 0.80, and 0.73, respectively, suggesting that this protein most likely binds metal. Another His residue, His95, is located nearby and likely contributes to metal binding. (*B*) Ni/Zn-binding site in the homology model of NYSGXRC-10060e.

seem to match likely metal-binding residues). Although the samples were of low concentration for EXAFS studies (<250 μM Zn concentrations), reasonable data could be collected to a k_{max} of 11 $Å^{-1}$. The Fourier-transformed EXAFS spectra for both target proteins indicated first shell average distances of ~2 Å, typical for Zn-binding (Fig. 4).

Active site models were created from crystal structure coordinates of similar proteins (PDB ID: 1M65, 19% identical to 9550a, and PDB ID: 2GWG, 60% identical to 9453d) such that simulations of the expected EXAFS spectra for the structure models could be generated and compared to the experimental data. The YcdX protein (PDB ID: 1M65, 19% identical to 9550a) appears to have three Zn-binding sites, one of which was considered to be high-affinity (Teplyakov et al. 2003). Target 9550a is remotely related to YcdX; however, the residues forming the three Zn sites are conserved. Protein from target 9550a was not specifically treated with high levels of Zn during purification. The HT-XAS-determined occupancy of 0.8 and presence of a high-affinity site in the 1M65 structure suggested that Zn is likely bound predominantly in a single site in this sample. A model of the Zn-binding site was created using the 1M65 coordinates, and EXAFS simulations for this structure were used to fit the first coordination shell EXAFS data for 9550a. The experimental data were found to be consistent with the overall metal–ligand distances and geometry of the high-affinity 1M65 model Zn-binding site (details of fit results for both targets are available in the Supplemental Material).

We examined the active site configuration in 2GWG (60% identical to 9453d) and determined that the bond lengths were inconsistent with typical Zn^{2+} binding. For example, the majority of the metal–N and metal–O distances were >2.4 Å, significantly longer than the ~2 Å canonical distances associated with these Zn–ligand bond lengths (Dimakis and Bunker 2004). Initial attempts to fit these Zn–ligand bond lengths to the EXAFS data were unsuccessful (data not shown). A model of the active site was created by modifying the existing 2GWG structure to decrease the Zn first shell distances, while maintaining the octahedral symmetry in 2GWG (Fig. 4). Simulated EXAFS data generated from this model were tested against the experimental EXAFS data and the best fit chosen as our representative Zn-binding site model. Two of the coordinating ligands are suggested to be oxygen atoms from nearby water molecules, located a bit further from the Zn (~2.5 Å) than the other ligands (~2.0 Å). Thus, in this case, the crystallized protein likely bound a nonnative metal atom which had long metal–ligand bond distances; Zn was incorrectly inserted in the map interpretation step. Our analysis correctly identified Zn as the native metal and accurately defined the first shell distances.

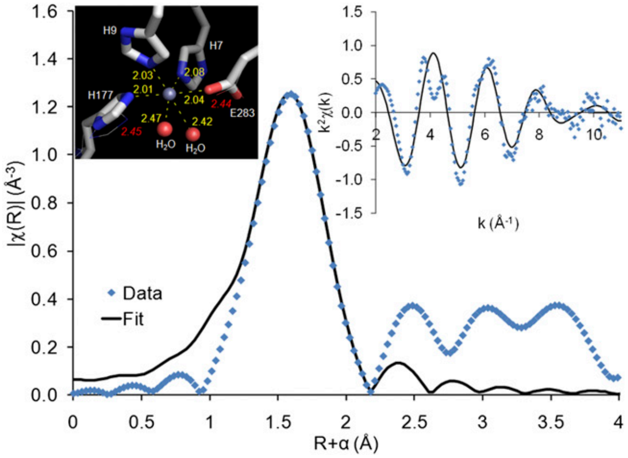

Figure 4. EXAFS data and model fit. Magnitude of the Fourier transform (uncorrected for phase shift) of the k^2-weighted 9453d EXAFS spectrum (blue) and first shell model fit (black). R-space fit performed from 1.3 to 2.4 Å. *Inset right*: k-space fit (2–11 $Å^{-1}$). *Inset left*: representation of 9453d Zn-binding site with distances labeled in yellow; original model based on 2GWG shown as thin lines (distances in red).

It is clear from these results that EXAFS studies of metalloproteomics targets, even at the most basic first shell analysis level, can provide important information for testing, if not validating, the proposed structures of the metal-binding sites. Even in cases where the crystal structure is solved for a target, the process of crystallization itself may impact the structure of the metal-binding site. An EXAFS scan of the target in its solution state allows relatively rapid recognition of such variation. For specified targets of interest, detailed EXAFS analysis and modeling can provide high-resolution three-dimensional structure information out to 5–6 Å from the bound metal (Levina et al. 2005).

Future directions

In the future, we plan to expand the sensitivity and range of the HT-XAS technology by installing superior solid-state detection technologies (with increased numbers of elements and increased data collection speed) and reconfiguring the X3B beamline to permit a wider range of elements to be analyzed, including selenium, molybdenium, and tungsten. Crystallographic phasing using Se-Met–labeled samples has become routine in protein crystallography. Our method could be used to confirm incorporation of Se into structure determination targets before crystallization. Molybdenum is an important trace element, since it is able to perform oxyanion catalysis without being too toxic. In most molybdenum-containing enzymes, the metal occurs as part of a specific molybdate cofactor, molybdopterin (Romao et al. 1995; Schindelin et al. 1996). Tungsten is the heaviest element with known biochemical/biological function. Tungsten-containing enzymes are found in eubacteria and archaea, mostly those with anaerobic metabolism (Andreesen and Ljungdahl 1973; Kletzin 1997). The L1-edge of tungsten occurs at ~12.1 keV, which is not far from the K-edge of selenium. A Si 220 crystal has been recently installed at beamline X3B at our facility at the National Synchrotron Light Source; this advance will permit access to the K-edge of molybdenum (20 keV). Although the HT-XAS method was initially developed as a new technology to take advantage of the large numbers of proteins being purified through structural genomics efforts, once we expand the capabilities of the beamline with greater sensitivity and range, we expect that the technology could be highly compatible with HT methods that fractionate biological samples or tissues and examine the fractions on an individual basis for metalloprotein detection and annotation and thus of general applicability for a wide range of metalloproteomic projects. In addition, our results indicate that MetalDetector has the potential to be a powerful method for genome-wide metal binding predictions, especially when a comprehensive training set is available.

Methods

Protein production

Protein samples were purified using the high-throughput protein production pipeline of the New York SGX Research Center for Structural Genomics (NYSGXRC) (Bonanno et al. 2005; Sauder et al. 2008). The majority of proteins were expressed in *E. coli* BL21 cells with C-terminal His_6 tags, although some were expressed with a cleavable N-terminal His_6-Smt3(SUMO) tag. Growth was typically in high yield selenomethionine (Se-Met) media (Orion Enterprises, Inc.), although LB or ZYP-5052 autoinduction media was sometimes used for native (non–Se-Met) protein production. Nearly all proteins were produced with the goal of crystallization, so the incorporation of selenium was to facilitate phasing for crystallographic structure determination. A subset of proteins (the amidohydrolases) was expressed with extra zinc in the presence of an iron chelator to minimize auto-oxidation of active site residues. Protein purification was performed using Ni affinity; fractions containing the protein were pooled and further purified by gel filtration chromatography on a GE Healthcare HiLoad 16/60 Superdex 200 prep grade column preequilibrated with gel filtration buffer (10 mM HEPES, pH 7.5, 150 mM NaCl, 10% glycerol, and 5 mM DTT). Proteins were typically concentrated to 5–10 mg/mL. SDS-PAGE and gel filtration chromatograms were stored for every protein, and proteins were required to pass a mass spectrometric quality control step. Matrix assisted laser desorption/ionization–time of flight (MALDI-TOF) and high performance liquid chromatography (HPLC)–electrospray ionization (ESI) mass spectrometry (MS) were used for intact mass analysis of all protein samples to examine their purity and chemical homogeneity and to compare measured molecular masses to the calculated molecular masses based on the theoretical protein sequences. The proteins were fermented in bacterial cells, so post-translational modifications were rare. Details of protein quality control by MS analysis are available in the Supplemental Material. Proteins that crystallized or had a sizeable mass discrepancy had their clones sequence-verified. Protein identification was performed for any potential reagent mix-ups, and sample information was corrected in the Laboratory Information Management System (LIMS), which served as a sample- tracking database. The information on the targets is included in the Supplemental Material (2006–2007_PSItarget.xls, 2008_PSItarget.xls, and 2009_PSItarget.xls) and can also be accessed through pepcDB (http://pepcdb.sbkb.org/).

Experimental setup

All experiments were done at beamline X3B of the National Synchrotron Light Source at Brookhaven National Laboratory. The experimental setup is designed to detect the following transition elements: Mn, Fe, Co, Cu, Ni, and Zn. The X-ray photons are produced by the NSLS X-ray ring, operating at a constant energy of 2.8 GeV with current decaying with time from 300 to 200 mA. The X-ray energy is set to 10 keV through a double-crystal monochromator that contains horizontal focusing optics. The apparatus consists of a multiplate, motorized rail, positioned at 45° with respect to the beam, that brings a sample plate into a position close to the synchrotron X-ray source and a 13-element, fast-count-rate, high-resolution Germanium detector placed perpendicular to the beam path (Supplemental Fig. S2). Each sample plate is a PTFE plate with 20 sample wells bored in. The appropriate volume of aqueous protein solution to obtain 0.1 mg of a protein is loaded in a sample well. After all samples are loaded, plates are placed in a hood, and samples are allowed to dry overnight. Samples were treated this way so that hundreds of samples can be loaded and measured quickly to enable the high-throughput nature of the method. The metal binding can be confirmed using fresh-frozen samples to collect a full edge scan (see EXAFS section) or corroborated with other computational methods discussed here. Standards and blank plates were prepared as described previously (Shi et al. 2005).

The synchrotron X-ray beam is concentrated using the sagittal focusing monochromator in the horizontal direction and further defined by slits to match the size of the sample well, 2.0 mm (vertical) × 6.5 mm (horizontal). Sample alignment and data recording are performed automatically by executing macros programmed using Labview software (National Instruments) on a computer running Windows XP. Since the Germanium detector simultaneously measures three energy regions (three metals), two separate runs are required to screen all six transition metals. Each sample is screened for the selected metals with sixty one-second-long counts by the 13-element Ge detector. The decay in X-ray current was taken into account, as the sum of all fluorescence signals was divided by the reference counts observed in the ion chamber.

Automation of the method

Previously, our experimental apparatus consisted of a small sample stage that could hold a maximum of 48 samples per experimental run. To optimize XAS data collection for a large number of samples with limited available beam time, a fast-moving motorized sample slide was constructed where a maximum of 11 sample plates could be simultaneously loaded onto a multiplate rail holding 220 samples. Data collection is controlled automatically by programmed macros. For a full HT-XAS experiment (220 samples), the total time to complete two runs to detect 6 metals is ~9 h, or ~2 1/2 min per sample. As we received ~120–150 samples each month from the NYSGXRC, one day of X3B beam time sufficed for the experimental setup and data collection.

We implemented several measures to reduce the noise and to improve the detection limit of the method. A vacuum beam pipe was installed to reduce the air absorption and air scatter to the detector. In addition, we improved the alignment precision between the X-ray beam and sample wells with the design of 20-well sample plates. These plates are thicker and more standardized, replacing the old 16-well plates that were easily distorted. A new automated protocol for X-ray beam setup and optimization, which slits the beam just under the size of the sample well to reduce the scatter from the X-ray beam hitting the sample plate, also decreased the noise significantly. With these measures, we were able to improve the limit of quantification (LOQ) by twofold, with the current LOQ of 0.5 μg of metal (20 nmol for a 12 kD protein).

Data analysis

A computer program written in the BASIC programming language was used to process the data and was described previously (Shi et al. 2005). Briefly, the program normalizes every timed count by dividing the detector channel with the beam current, adds 13 detector channels, and subtracts background counts measured for the blank plate. For every sample plate, the output file contains the resulting counts for all metals in each sample well. In addition, the program applies statistical functions for better peak discrimination. An average fluorescence count (F_{ave}) is computed from fluorescence counts (F) of all 6 metals in 20 samples (from a single sample plate). Samples with $F/F_{\text{ave}} > 5$ are further investigated for metal binding.

Homology modeling and metal binding prediction

Most (>80%) of the proteins purified through the structural genomics pipelines do not yield high-resolution structures; however, homology models are available for some of these targets in MODBASE (Pieper et al. 2004). Others were obtained using the SWISS-MODEL (Kiefer et al. 2009) server, fed with alignments from PSI-BLAST.

Metal binding was predicted using MetalDetector (Passerini et al. 2006; Lippi et al. 2008). The program combines two machine-learning–based bonding state predictors (specialized on metal ions and disulfide bonds, respectively), both trained on carefully curated data sets. In particular, the metal-binding data set consists of 2727 nonredundant protein chains extracted from PDB, 365 of which contain at least one metal-binding site. Redundancy was removed by running UniqueProt (Mika and Rost 2003), ensuring that no pair of proteins had a positive HSSP value. Metal-bonding state of every Cys and His was determined by parsing PDB files and searching for transition metals and transition metal complexes. Any residue having a heavy atom within 3.0 Å of the metal (or complex) was labeled as a ligand. In heme and Fe/S complexes, residues binding to the porphyrin ring and to the sulfur atoms were also considered as ligands. Low-resolution PDB entries and chains containing a single metal-binding residue were not included in the data set. These criteria were motivated by the need to obtain a large and representative data set and, at the same time, to avoid including training examples that were unlikely to be functionally relevant.

MetalDetector features two user-adjustable thresholds, T_D and T_M, that allow controlling the trade-off between precision and recall for the disulfide-bound and the metal-bound state, respectively. In particular, T_D and T_M are thresholds to the probabilities of Cys and His residues in the disulfide-bridged state (applied only to Cys) or in the metal-bound state. For assessing the prediction overlap between HT-XAS and MetalDetector, we used MetalDetector in a high metal recall mode (Lippi et al. 2008) by adjusting $T_D = 1$ and analyzing values of T_M between 0 and 1. Lower T_M values correspond to a higher chance of predicting a residue as metal-bound. To reduce bias, we created a redundancy-reduced data set. Indeed, several proteins in the set of nearly 4000 protein samples analyzed by HT-XAS (3431 distinct sequences, of which 343 were identified by HT-XAS as metalloproteins) shared rather high levels of sequence similarity. Thus, for the comparison between experimental and computational data, we considered a sequence redundancy threshold of 70% (using the program CD-HIT) (Li and Godzik 2006). This gave a total of 2453 sequences; 42 sequences that had no His and no Cys (i.e., cases that MetalDetector cannot predict as metal-binding) were further removed, giving a total of 2411; 234 of these were identified as metalloproteins by HT-XAS (i.e., slightly less than 10%). We note that ~42% of the proteins analyzed here can be aligned by BLAST with an E-value < 10^{-3} with a protein in the MetalDetector training set (Passerini et al. 2006). Removing these sequences from our data set does not seem to have a strong effect on MetalDetector performance relative to HT-XAS assignments. When removing these sequences, at 10% recall, precision is 36 ± 5% and 67 ± 7% for $N = 1$ and $N = 5$, respectively, compared with 42 ± 4% and 66 ± 5% for the entire set. Break-even points for $N = 1$ and $N = 5$ are 32% and 41%, respectively, compared with 32% and 45% on the entire set.

MetalDetector predictions for the entire HT-XAS set and a catenated fasta file of all sequences in the redundancy reduced set can be accessed at http://metaldetector.dsi.unifi.it/HT-XAS.zip.

To quantify agreement between HT-XAS and MetalDetector, we defined two quantities: relative-precision (TP*/[TP*+FP*]) and relative-recall (TP*/[TP*+FN*]). These quantities correspond to the usual precision (aka positive predictive value) and recall (aka sensitivity) measures when the standard of truth for determining whether or not a target is a metalloprotein is the HT-XAS assignment. In other words, TP* is the number of metalloproteins predicted by MetalDetector that are identified as such by HT-XAS (ratio ≥ 0.3), while FP* is the number of metalloproteins predicted by MetalDetector that are not identified as such by HT-XAS; FN* is the number of targets that MetalDetector does not predict as metalloproteins while HT-XAS does.

For both relative-precision and relative-recall, we calculated averages and standard errors by performing bootstrapping without resampling (Efron and Tibshirani 1993); for this purpose, we created 100 sets comprised of 1808 randomly selected proteins (i.e., 75% of 2411).

Extended X-ray absorption fine structure (EXAFS)

EXAFS data were collected on 9550a and 9453d samples at the NSLS X3B beamline as K_α fluorescence spectra. For each sample, 26 scans, with 8–11 Ge fluorescence detector channels averaged for each scan, were taken over a range from 200 eV below to 16 × k above the Zn K-edge (defined as 9659 eV). Data processing and first shell analysis were performed using the IFEFFIT software (Newville 2001; Ravel and Newville 2005). Details of EXAFS data collection and analysis are available in the supplemental information. The raw EXAFS data for the two targets are also attached as Supplemental Materials (9453d_EXAFS.txt and 9550a_EXAFS.txt).

Acknowledgments

This work is partly supported by NIH research grant U54 GM074945 (PI: S.K.B.). Data for this study were collected at beamline X3B of the National Synchrotron Light Source, Brookhaven National Laboratory. Financial support for the operation of the beamline comes principally from NIBIB, under grant P30-EB-09998, and the Office of Basic Energy Sciences of the U.S. Department of Energy (DE-AC02-98CH10886). M.P. would like to acknowledge the support of the Technische Universität München, Institute for Advanced Study, funded by the German Excellence Initiative.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402.

Andreesen JR, Ljungdahl LG. 1973. Formate dehydrogenase of *Clostridium thermoaceticum*: Incorporation of selenium-75, and the effects of selenite, molybdate, and tungstate on the enzyme. *J Bacteriol* **116:** 867–873.

Andreini C, Banci L, Bertini I, Rosato A. 2006. Zinc through the three domains of life. *J Proteome Res* **5:** 3173–3178.

Andreini C, Bertini I, Rosato A. 2009. Metalloproteomes: A bioinformatic approach. *Acc Chem Res* **42:** 1471–1479.

Babbitt PC, Hasson MS, Wedekind JE, Palmer DR, Barrett WC, Reed GH, Rayment I, Ringe D, Kenyon GL, Gerlt JA. 1996. The enolase superfamily: A general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. *Biochemistry* **35:** 16489–16501.

Banci L, Bertini I, Ciofi-Baffoni S, Gerothanassis IP, Leontari I, Martinelli M, Wang S. 2007. A structural characterization of human SCO2. *Structure* **15:** 1132–1140.

Bartel J, Charkiewicz E, Bartz T, Schmidt D, Grbavac I, Kyriakopoulos A. 2010. Metalloproteome of the prostate: Carcinoma cell line DU-145 in comparison to healthy rat tissue. *Cancer Genomics Proteomics* **7:** 81–86.

Bernstein DA, Zittel MC, Keck JL. 2003. High-resolution structure of the *E. coli* RecQ helicase catalytic core. *EMBO J* **22:** 4910–4921.

Bertini I, Cavallaro G. 2008. Metals in the "omics" world:Copper homeostasis and cytochrome c oxidase assembly in a new light. *J Biol Inorg Chem* **13:** 3–14.

Bettmer J. 2005. Metalloproteomics: A challenge for analytical chemists. *Anal Bioanal Chem* **383:** 370–371.

Bonanno JB, Almo SC, Bresnick A, Chance MR, Fiser A, Swaminathan S, Jiang J, Studier FW, Shapiro L, Lima CD, et al. 2005. New York-Structural GenomiX Research Consortium (NYSGXRC): A large scale center for the protein structure initiative. *J Struct Funct Genomics* **6:** 225–232.

Chance MR, Fiser A, Sali A, Pieper U, Eswar N, Xu G, Fajardo JE, Radhakannan T, Marinkovic N. 2004. High-throughput computational and experimental techniques in structural genomics. *Genome Res* **14:** 2145–2154.

Cvetkovic A, Menon AL, Thorgersen MP, Scott JW, Poole FL II, Jenney FE Jr, Lancaster WA, Praissman JL, Shanmukh S, Vaccaro BJ, et al. 2010. Microbial metalloproteomes are largely uncharacterized. *Nature* **466:** 779–782.

Degtyarenko K. 2000. Bioinorganic motifs: Towards functional classification of metalloproteins. *Bioinformatics* **16:** 851–864.

Dimakis N, Bunker G. 2004. XAFS Debye-Waller factors for Zn metalloproteins. *Phys Rev B* **70:** 195114. doi: 10.1103/PhysRevB.70.195114.

Efron B, Tibshirani RJ. 1993. *An introduction to bootstrap*. Chapman and Hall, New York.

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, et al. 2008. The Pfam protein families database. *Nucleic Acids Res* **36:** D281–D288.

Graslund S, Nordlund P, Weigelt J, Hallberg BM, Bray J, Gileadi O, Knapp S, Oppermann U, Arrowsmith C, Hui R, et al. 2008. Protein production and purification. *Nat Methods* **5:** 135–146.

Humphrey W, Dalke A, Schulten K. 1996. VMD: Visual molecular dynamics. *J Mol Graph* **14:** 33–38.

Kiefer F, Arnold K, Kunzli M, Bordoli L, Schwede T. 2009. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res* **37:** D387–D392.

Kletzin A. 1997. *Tungsten-containing aldehyde ferredoxine oxidoreductase*. OPA, Amsterdam, Netherlands.

Kulkarni PP, She YM, Smith SD, Roberts EA, Sarkar B. 2006. Proteomics of metal transport and metal-associated diseases. *Chemistry* **12:** 2410–2422.

Levina A, Armstrong RS, Lay PA. 2005. Three-dimensional structure determination using multiple-scattering analysis and XAFS: Applications to metalloproteins and coordination chemistry. *Coord Chem Rev* **249:** 141–160.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22:** 1658–1659.

Lippi M, Passerini A, Punta M, Rost B, Frasconi P. 2008. MetalDetector: A web server for predicting metal-binding sites and disulfide bridges in proteins from sequence. *Bioinformatics* **24:** 2094–2095.

Manjasetty BA, Shi W, Zhan C, Fiser A, Chance MR. 2007. A high-throughput approach to protein structure analysis. *Genet Eng (NY)* **28:** 105–128.

Manley SA, Byrns S, Lyon AW, Brown P, Gailer J. 2009. Simultaneous Cu-, Fe-, and Zn-specific detection of metalloproteins contained in rabbit plasma by size-exclusion chromatography-inductively coupled plasma atomic emission spectroscopy. *J Biol Inorg Chem* **14:** 61–74.

Mika S, Rost B. 2003. UniqueProt: Creating representative protein sequence sets. *Nucleic Acids Res* **31:** 3789–3791.

Min J, Landry J, Sternglanz R, Xu RM. 2001. Crystal structure of a SIR2 homolog-NAD complex. *Cell* **105:** 269–279.

Newville M. 2001. EXAFS analysis using FEFF and FEFFIT. *J Synchrotron Radiat* **8:** 96–100.

Oyama T, Oka H, Mayanagi K, Shirai T, Matoba K, Fujikane R, Ishino Y, Morikawa K. 2009. Atomic structures and functional implications of the archaeal RecQ-like helicase Hjm. *BMC Struct Biol* **9:** 2. doi: 10.1186/1472-6807-9-2.

Passerini A, Punta M, Ceroni A, Rost B, Frasconi P. 2006. Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. *Proteins* **65:** 305–316.

Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP, Stuart AC, Mirkovic N, Rossi A, Marti-Renom MA, Fiser A, et al. 2004. MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res* **32:** D217–D222.

Pieper U, Chiang R, Seffernick JJ, Brown SD, Glasner ME, Kelly L, Eswar N, Sauder JM, Bonanno JB, Swaminathan S, et al. 2009. Target selection and annotation for the structural genomics of the amidohydrolase and enolase superfamilies. *J Struct Funct Genomics* **10:** 107–125.

Pike AC, Shrestha B, Popuri V, Burgess-Brown N, Muzzolini L, Costantini S, Vindigni A, Gileadi O. 2009. Structure of the human RECQ1 helicase reveals a putative strand-separation pin. *Proc Natl Acad Sci* **106:** 1039–1044.

Ravel B, Newville M. 2005. ATHENA, ARTEMIS, HEPHAESTUS: Data analysis for X-ray absorption spectroscopy using IFEFFIT. *J Synchrotron Radiat* **12:** 537–541.

Romao MJ, Archer M, Moura I, Moura JJ, LeGall J, Engh R, Schneider M, Hof P, Huber R. 1995. Crystal structure of the xanthine oxidase-related aldehyde oxido-reductase from D. gigas. *Science* **270:** 1170–1176.

Sauder JM, Rutter ME, Bain K, Rooney I, Gheyi T, Atwell S, Thompson DA, Emtage S, Burley SK. 2008. High throughput protein production and crystallization at NYSGXRC. In *Methods in molecular biology: Structural proteomics: High-throughput methods* (ed. B Kobe, M Guss, T Huber), Vol. 426, pp. 561–575. Humana, Totowa, NJ.

Schindelin H, Kisker C, Hilton J, Rajagopalan KV, Rees DC. 1996. Crystal structure of DMSO reductase: Redox-linked changes in molybdopterin coordination. *Science* **272:** 1615–1621.

Scott RA, Shokes JE, Cosper NJ, Jenney FE, Adams MW. 2005. Bottlenecks and roadblocks in high-throughput XAS for structural genomics. *J Synchrotron Radiat* **12:** 19–22.

Shi W, Chance MR. 2006. Structural genomics: High-throughput structure determination of protein domains. In *Comprehensive medicinal chemistry* (ed. DJ Triggle, JB Taylor), Vol. 3. Elsevier, Amsterdam, Netherlands.

Shi W, Chance MR. 2008. Metallomics and metalloproteomics. *Cell Mol Life Sci* **65:** 3040–3048.

Shi W, Zhan C, Ignatov A, Manjasetty BA, Marinkovic N, Sullivan M, Huang R, Chance MR. 2005. Metalloproteomics: High-throughput structural and functional annotation of proteins in structural genomics. *Structure* **13:** 1473–1486.

Shu N, Zhou T, Hovmoller S. 2008. Prediction of zinc-binding sites in proteins from sequence. *Bioinformatics* **24:** 775–782.

Szpunar J. 2005. Advances in analytical methodology for bioinorganic speciation analysis: Metallomics, metalloproteomics, and heteroatom-tagged proteomics and metabolomics. *Analyst (Lond)* **130:** 442–465.

Teplyakov A, Obmolova G, Khil PP, Howard AJ, Camerini-Otero RD, Gilliland GL. 2003. Crystal structure of the *Escherichia coli* YcdX protein reveals a trinuclear zinc active site. *Proteins* **51:** 315–318.

Yew WS, Fedorov AA, Fedorov EV, Wood BM, Almo SC, Gerlt JA. 2006. Evolution of enzymatic activities in the enolase superfamily: D-tartrate dehydratase from *Bradyrhizobium japonicum*. *Biochemistry* **45:** 14598–14608.

Received September 9, 2010; accepted in revised form March 4, 2011.