research articles Journal of proteome research

Occurrence of Copper Proteins through the Three Domains of Life: A Bioinformatic Approach

Claudia Andreini, Lucia Banci, Ivano Bertini,* and Antonio Rosato

Magnetic Resonance Center (CERM) and Department of Chemistry, University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy

Received July 30, 2007

In high-throughput genome-level protein investigation efforts, such as Structural Genomics, the systematic experimental characterization of metal-binding properties (i.e., the investigation of the metalloproteome) is not always pursued and remains far from trivial. In the present work, we have applied a bioinformatic approach to investigate the occurrence of (putative) copper-binding proteins in 57 different organisms spanning the entire tree of life. We found that the size of the copper proteome is generally less than 1% of the total proteome of an organism, in both eukaryotes and prokaryotes. The occurrence of copper-binding proteins is relatively scarce when compared to that of zinc-binding proteins and of non-heme iron proteins. This may be due to both poorer bioavailability (in particular with respect to iron in the ancient world) and the complexity of copper chemistry and the risks associated with it, which may have adversely affected natural selection of copper-binding proteins. The present analysis shows that there is a strong relationship between the metal coordination sphere and protein function. A network involving proteins having roles in both copper transport and respiration was identified, parts or all of which are detected in the majority of the organisms examined.

Keywords: copper • metalloprotein • metalloproteome • evolution • electron transfer • homeostasis

Introduction

Copper is an essential trace metal utilized as a cofactor in a variety of proteins. In eukaryotes, copper-dependent metalloenzymes are found in multiple cellular locations.[1] Excess copper, however, is highly toxic to most organisms.[1,2] Accordingly, a complex machinery of proteins that bind the metal ion controls the uptake, transport, sequestration, and efflux of copper in vivo.[3–5] Indeed, the concentration of free copper ions should be maintained at an essentially negligible level, as the copper ions may catalyze the formation of radicals which can damage cell membranes. On the other hand, newly produced copper-binding proteins need to uptake copper ions to achieve their mature, active form. This dual goal can be obtained if systems permitting rapid and efficient metal transfer and simultaneously preventing nonspecific reactions involving copper are in place. In particular, so-called metallochaperones, which deliver copper to specific intracellular targets, lower the activation barrier for copper transfer to their specific partners,[6] thereby circumventing the significant thermodynamic overcapacity for copper chelation of the cytoplasm.[7] As an example, one of the best studied pathways of copper transfer present in humans involves a small soluble metallochaperone, HAH1 (also known as Atox1),[8,9] which is capable of delivering copper(I) to two membrane-bound P-type ATPases, called, respectively, the Menkes and the Wilson proteins (ATP7A and ATP7B, respectively).[3–5] These, in turn, can translocate copper in the trans-Golgi network or across the plasma membrane,[3–5] depending on environmental conditions.[10] NMR spectroscopy and X-ray crystallography have provided many structures of different copper chaperones from a variety of different organisms, which have afforded many interesting details on the atomic-level mechanisms for copper control and transfer.[11,12]

In 2004, we endeavored to devise a method to exploit information present in the PDB[13] for mining genome sequences (or gene databanks) to identify metalloproteins, using copper-binding proteins as a test case.[14] The approach was based on the analysis of the occurrence of conserved patterns of amino acids that were known (from the analysis of the PDB) to bind a metal ion, together with sequence similarity requirements. In short, a library of copper-binding patterns (hereafter CBP) was automatically built from all copper-binding proteins in the PDB.[13,15] Each CBP was then used together with the primary sequence of the corresponding metalloprotein to browse any ensemble of sequences of interest with the program PHI-BLAST,[16] using the level of sequence similarity around the CBP as a filter criterion. The proteins in the PDB were used to assess the search procedure in terms of potentiality and quality of the output.[14] In that very same work, we realized that the procedure outlined above retrieved a significant number of metalloproteins which were known to bind metals other than copper to carry out their physiological function. For some CBP types, the noncopper metalloproteins retrieved actually outnumbered the true copper-binding proteins identified.[14] To correct this bias, while at the same time retaining the good sensitivity of the above approach, we integrated searches based

* Corresponding author. Magnetic Resonance Center, University of Florence, Via L. Sacconi, 6 50019 Sesto Fiorentino, Italy. Fax: +39 055 4574271. Tel: +39 055 4574272. E-mail: ivanobertini@cerm.unifi.it.

10.1021/pr070480u CCC: $40.75

Published on Web 11/08/2007

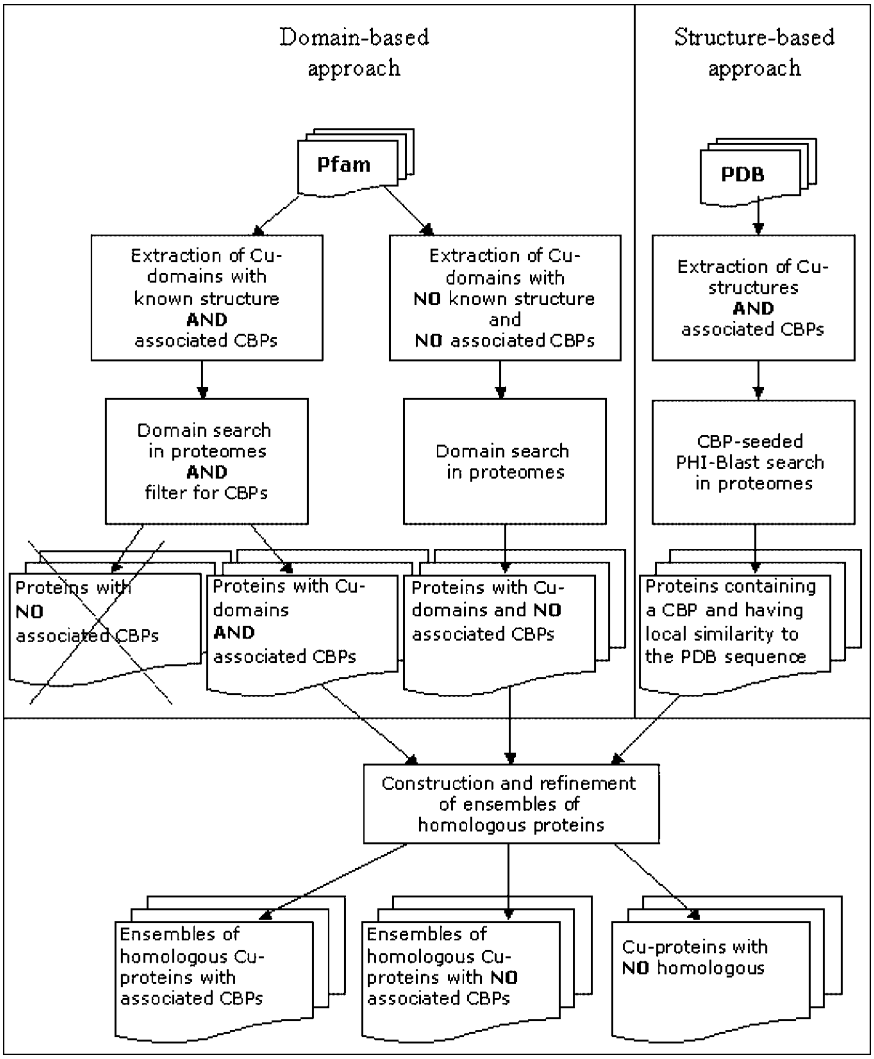
Domain-based approach
Structure-based approach
Pfam
PDB
Extraction of Cu-domains with known structure AND associated CBPs
Extraction of Cu-domains with NO known structure and NO associated CBPs
Extraction of Cu-structures AND associated CBPs
Domain search in proteomes AND filter for CBPs
Domain search in proteomes
CBP-seeded PHI-Blast search in proteomes
Proteins with NO associated CBPs
Proteins with Cu-domains AND associated CBPs
Proteins with Cu-domains and NO associated CBPs
Proteins containing a CBP and having local similarity to the PDB sequence
Construction and refinement of ensembles of homologous proteins
Ensembles of homologous Cu-proteins with associated CBPs
Ensembles of homologous Cu-proteins with NO associated CBPs
Cu-proteins with NO homologous

Figure 1. Scheme of the protocol used for the identification of copper-binding proteins.

on metal-binding patterns with domain recognition methods, obtaining quite satisfactory results in terms of both sensitivity and selectivity. These results were already published for zinc-binding and nonheme iron-binding proteins.[17–19] The results for human zinc proteins proved quite solid also when compared to completely independent predictions based on machine learning methods.[20,21] This methodological advancement prompted us to re-examine the occurrence of copper-binding proteins in completely sequenced genomes using a significantly larger number of organisms than previously employed.[14] The results of this new analysis are discussed in the present work. The consistent application of an advanced computational approach for different metals on a common set of organisms will enable larger-scale comparative analyses of metalloproteomes.

Methods

The protocol used to retrieve copper-binding proteins in the proteomes is depicted in Figure 1 and is described below. A detailed graphical scheme of the protocol is also given in Supplementary Figure S1 (Supporting Information). Figure S1A describes the preparation of input files, and Figure S1B describes the production of output (i.e., the retrieval of putative copper-binding proteins).

1. Preparation of Input Files. All the structures deposited in the PDB[13,15] as of January 2007 and containing at least one copper ion in the coordinate file were extracted. Only one representative was kept for all proteins binding the metal ion(s) with the same pattern(s) and having a sequence identity greater than 98%. Proteins binding copper in a nonphysiologically relevant manner (e.g., copper-substituted structures like the 2oxi structure[22]) were manually removed. The resulting representative ensemble contained 164 copper-binding structures. The coordinate files of each structure were used to identify the residues coordinating the metal ion(s), which define the CBP, with an approach already applied to other classes of metalloproteins.[17–19] Every residue having at least one heavy atom at a distance shorter than 2.8 Å from the metal was defined as a metal–ligand. Dimetallic clusters were considered as a single center, and thus all the ligands to the various copper ions formed a single CBP. Note that more than one CBP can be associated to a single protein, depending on the number of metal ions bound. As a result of this analysis, a library of 157 distinct CBPs was assembled (Table S1 in the Supporting Information).

In parallel, we extracted copper-binding protein domains from the Pfam library of domains.[23] All domains whose description contained the words “Copper” or “Cu” have been selected and manually analyzed to collect a list of copper-binding domains constituting the input for searches in each proteome. The ensemble of the retrieved domains was then manually refined removing false positives and adding domains that, although not annotated as copper-binding in Pfam, have been reported in the literature as such (e.g., Cox19[24] and serum albumin[25]). Domains whose physiological association with copper has been suggested but is not supported in the recent literature have been excluded (e.g., hemoglobin[26,27]). To achieve maximal coverage, copper-binding protein structures retrieved from the PDB[13] have been analyzed in terms of their domain content to identify domains always or commonly containing a copper-binding site but not annotated as such in Pfam.[23] This procedure also allowed us to associate CBPs to Pfam[23] domains having at least one structurally characterized representative. In this way, we created an ensemble of copper-binding domains, most of which have an associated CBP (Table S2 in the Supporting Information).

2. Proteomes Selection and Search. Proteome sequences for the selected organisms (Table S3 in the Supporting Information) have been downloaded from the Entrez Genome (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome) Web site at the National Center for Biotechnology Information. The organisms have been selected to map evenly the tree of life (Figure S2 in the Supporting Information). We selected archaeal and bacterial organisms by taking one organism per order; when there was more than one organism sequenced in a order, we selected the organism sequenced first. 40 bacterial, 12 archaeal, and 5 eukaryotic proteomes were analyzed.

Each of the 57 selected proteomes was then searched for protein sequences containing a copper-binding domain with the program HMMER,[28] using as input the HMMs available from Pfam.[23] For copper-binding domains with an associated CBP (or CBPs), we filtered the results by discarding the sequences containing the domain but lacking the CBP (Figure 1). Some variability was allowed for the CBP, by letting the spacing between the metal and ligands vary within ±20% (or by ±1 residue when the spacing was less than 5 residues). As a complementary analysis, we ran PHI-BLAST[16] searches on the same proteomes using each CBP as a seed and the sequence of the structure containing it as a query (Figure 1). Hits with a fraction of identical amino acids aligned by PHI-BLAST[16] with respect to the length of the query protein (I_d^{global}) greater than 0.20, for CBPs with at least three ligands, or with $I_d^{global} > 0.30$, for CBPs of only two ligands, were kept as putative copper-binding proteins. The latter approach has been applied in previous studies.[17–19]

All putative copper-binding proteins retrieved were pooled together and grouped based on homology (Figure 1). Homologous proteins were identified by performing BLAST[29] searches for each putative copper-binding protein against all the other identified copper-binding proteins. Proteins were grouped in a way such that each ensemble contained all the hits within a given threshold (*E*-value calculated by BLAST $< 10^{-3}$) for each of its members. An additional restraint imposed was that the length of the protein sequences had to be within a 2-fold. All ensembles of sequences were manually inspected to check whether two (or more) different protein families (e.g., different domain composition, different CBP, etc.) had been grouped together. Then, a profile was built for each ensemble of sequences and used to further search the proteomes to extend our coverage. Each profile thus describes a group of proteins similar in sequence and with a similar CBP. Newly identified proteins were then filtered on the basis of the presence of the CBP and subsequently added to the ensemble corresponding to the profile. The latter was recalculated, and profile searches were repeated until no new sequences could be identified. At the end of the procedure described in this section, a database of putative copper-binding protein sequences was obtained. The implemented procedure produced ensembles of proteins sharing sequence similarity. The majority of ensembles also had an associated CBP, which was conserved with little variation in spacing in all proteins belonging to each ensemble.

3. Data Analysis. To obtain functional hints, all sequences retrieved were analyzed against the entire Pfam domain library[23] as well as in terms of their gene ontology (GO) functional annotation.[30] To study potential functional associations among the ensembles of copper proteins having representatives in all the domains of life, a COG identification number (Cluster of Orthologs Group)[31] was associated to all these ensembles. The list of COG identification numbers was then used to query the STRING database,[32,33] using the *COG mode* function and the *multiple names* interface. The STRING database was searched using as prediction criteria all those available in the database interface except for (i) databases and (ii) text mining and requiring high-confidence scores (≥0.7). The subcellular localization of the potential copper proteins was predicted using PSORTb[34,35] for the prokaryotic proteins and WoLF PSORT[36] for the eukaryotic proteins. The prediction of the subcellular localization of archaeal sequences was performed using the Gram-positive option of PSORTb. When available, the subcellular localization of the human proteins was taken from LOCATE,[37] a database containing subcellular localization data derived from high-throughput, immunofluorescence-based assays and the literature. We observed that for about 50% of the human proteins the subcellular localization predicted by WoLF PSORT[36] was different from that reported in LOCATE.[37]

4. Evaluation of Methods Performance. To evaluate the performance of our protocol, we applied it to the entire PDB (thus, including both copper proteins and all other proteins) and computed performance measures. These measures included precision and recall. For a given target class C (e.g., copper binding proteins), let true positives (TP) be the number of examples correctly predicted as belonging to C, false positives (FP) be the number of examples incorrectly predicted as belonging to C, and false negatives (FN) be the number of examples of class C incorrectly assigned to another class. Precision is the number of true positives over the total number of examples assigned to the class, that is, TP/(TP + FP). Recall (or coverage) is the number of true positives over the total number of examples belonging to the class, that is, TP/(TP + FN). In the case of binary classification (i.e., yes/no as in the present case), recall for the positive class is also known as sensitivity.

The ensemble of positives was built taking from the PDB[13,15] one representative sequence for each group of copper-binding proteins having a sequence identity greater than 90% (artifacts were manually removed). The ensemble of negatives was built similarly, taking one representative for all other proteins in the PDB. These two data sets were searched using the input files prepared as detailed in the previous sections. The performance of our approach was evaluated for different HMMER[28] *E*-values and different I_d^{global} thresholds. For PHI-BLAST searches,[16] we removed from the results all hits to the query itself (which is a protein of known structure) and to its homologues so as to not

Table 1. Percentage of CBPs Containing S-, N-, and O-Donors for Different Classes of Copper-Binding Sites

	homeostasis	redox centers (copper cycles between different oxidation states)				
	Cu(I) sites	type I	type II	type III	CuA	CuB
S-donors (i.e., Met, Cys)	100%	89%	9%	0%	100%	0%
N-donors (i.e., His)	0%	100%	100%	100%	100%	100%
O-donors (i.e., Asp, Glu, Tyr)	0%	7%	21%	0%	45%	0%

introduce a favorable bias in the performance measures, as previously described.[14]

Results and Discussion

1. Input Preparation and Evaluation of Performance. 475 protein structures containing at least one copper ion were retrieved from the PDB,[13] corresponding to 164 nonredundant copper-binding protein sequences and 157 distinct CBPs (Table S1, Supporting Information). 39 Pfam domains[23] were identified as copper binding (Table S2, Supporting Information). These data were used as the starting point to search in the selected proteomes. When the search was applied to the PDB[13] to evaluate the performance of our protocol, the structure-based approach alone had a recall (fraction of copper proteins identified with respect to all copper proteins in the PDB[13]) of 0.69 with a precision of 0.87 (fraction of proteins correctly predicted to be copper binding). At a HMMER *E*-value threshold[28] of 0.05, the domain-based searches filtered for the CBPs had a recall of 0.96 and a precision of 0.83. If no filtering was applied after Pfam[23] searches, the recall remained unchanged, as expected, while the precision dropped to 0.73. Pfam domain[23] searches retrieved more true copper proteins than PHI-BLAST,[16] hence the higher recall, but also, the number of false positives retrieved was, even after filtering, slightly larger.

2. Searching for Copper-Binding Proteins in the Proteomes. We identified 401, 119, and 396 distinct putative copper-binding proteins in 40 bacterial, 12 archaeal, and 5 eukaryotic organisms, respectively (Tables S4–6, Supporting Information), including 304 sequences from domains without any structure available (33% of all sequences retrieved). Indeed, many domains are known to be involved in copper homeostasis (like NosD) but either have unknown structures or only the apo form structure is available, so it was not possible to associate a CBP to them.

The number of copper ligands contained in a single CBP was found to vary between two and six, with three being the most common coordination number. Note however that, because dimetallic clusters are treated as a single block, this number does not always reflect the coordination geometry around each individual copper ion but is actually correctly defined as the number of amino acids involved in binding the cofactor. In fact, all the examples of patterns having six ligands are involved in the coordination of dinuclear copper centers (e.g., CuA). The most common protein–ligands for copper are histidines and cysteines, both found in about 65% of the patterns retrieved. Another common ligand for copper is methionine, which is found in about 15% of the cases. Glutamate and aspartate as well as tyrosine are less common (about 7%, 6%, and 1%, respectively). In biological systems, copper ions (not taking into account polymetallic clusters) can have two oxidation states, i.e., +1 and +2, which have quite different coordination chemistry. Copper(I) is relatively soft and prefers S-donor ligands, whereas copper(II) is relatively hard and prefers N-donor ligands.[38]

A more detailed analysis of the retrieved patterns, performed classifying copper proteins in functional groups, highlights that there is a strong correlation between the CBP nature and copper protein function (Table 1) and, consequently, its oxidation state. In particular, the large majority of the proteins involved in copper transport and homeostasis binds copper(I) (Figure S4, Supporting Information) and has CBPs containing only sulfur donor atoms from cysteine and methionine side chains. A possible exception to this general behavior is given by Sco proteins, which may involve a histidine residue in copper coordination together with two cysteines. This may be associated with the fact that Sco proteins are able to bind both copper(I) and copper(II).[39–41a] Copper(I) sites also tend to have low coordination numbers, typically two with the notable exception of polycopper clusters such as in COX2 or in metallothionein. The selection of the coordination number two for proteins involved in processes such as copper homeostasis and transport can be explained because copper binding in these processes is transient and proteins must have the possibility to exchange the ion among them. The availability of unfilled coordination positions allows the coordination number of copper to transiently increase in the course of metal transfer between two partner proteins.[41b] The CBPs of most copper(I) transporters are closely spaced, with only two residues between the two copper-binding cysteines.

Owing to its redox properties, nature has widely exploited copper as the catalytic site of enzymes catalyzing redox reactions or as the redox site in electron transfer proteins. In these systems, the copper ion cycles between the +1 and +2 oxidation states (again neglecting polycopper clusters).[38] Therefore, copper enzymes must have a coordination sphere for the metal capable of accommodating it in both oxidation states. Tuning of the coordination sphere in fact may constitute a suitable means to tune the reduction potential of the metal itself. Historically, copper sites where copper behaves as a redox center were classified into three different types on the basis of their spectroscopic properties (type I, type II, and type III) (Figure S4, Supporting Information). In fact, the spectroscopic properties reflect the copper coordination sphere so that these three types of coordination spheres represent operative solutions to the accommodation of copper(I) and copper(II) in the same site. Note that we have detected proteins that are similar in sequence to type I copper proteins but for which we could not identify all metal ligands on the basis of the presence of a CBP already contained in a type I protein of known structure. In these cases, we have identified at least a pair of His residues with spacing similar to that of a known CBP and checked that the sequence contained also Cys residues. To reflect this, we did not assign the putative Cys ligand, and therefore the percentage of sites containing S-donors in Table 1 is less than 100% even though the presence of a Cys residue is an intrinsic characteristic of a type I CBPs. Other well-known redox copper sites are that of cytochrome *c* oxidase (CuA and CuB). These sites again are another example of tuning the redox potential

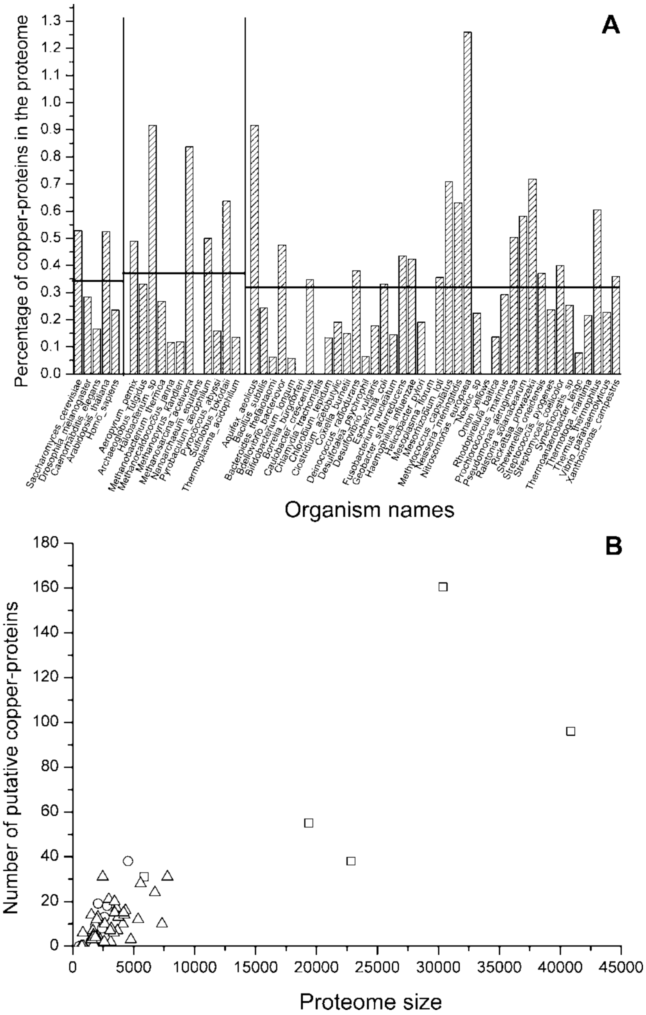

Figure 2. (A) Percentage of copper-binding proteins identified in each proteome analyzed. Each number on the *x*-axis corresponds to one organism, as reported in Table S3. Vertical lines separate the three domains. Horizontal lines show the average values within each domain (eukaryota, archaea, and bacteria). (B) Number of putative copper-binding proteins as a function of the proteome size (squares, eukaryota; circles, archaea; triangles, bacteria).

of copper, particularly for the CuA cofactor whose dicopper site can cycle between +3 and +2. These CBPs always contain more than two ligands, which is a requisite for stable binding of the copper(II) ion. Histidine is the most common ligand in these sites, and indeed some CBPs are made up by only histidine side chains. However, it is not uncommon that copper-containing enzymes or electron transfer proteins have CBPs that are constituted by a mixture of N-, O-, and S-donors (e.g., in blue copper proteins). From our analysis, it appears that no protein binds exclusively copper(II). All known and well-studied copper-binding proteins are either copper transporters, which bind copper(I) almost invariably, or redox enzymes and electron-transport proteins in which copper cycles between copper(I) and copper(II).

3. Distribution of Copper Proteins in the Three Domains of Life. Figure 2A reports the percentage of copper-binding proteins in each of the proteomes analyzed. It can be seen that the share of copper proteins is quite constant across the various proteomes from bacteria to humans, representing less than 1% of the proteome. Bacterial proteomes contain between 0% and 1.3% copper proteins, with an average of 0.3% (from 0 to 31 proteins). Archaeal proteomes contain between 0% and 0.9% copper proteins, with an average of 0.4% (from 0 to 38 proteins). Finally, eukaryotic proteomes contain between 0.2% and 0.5%, with an average of 0.3% (from 31 to 160 proteins). The data are not biased by the original selection of structures from the PDB, as 64% are bacterial and 36% are eukaryotic, whereas only a negligible fraction of structures in the PDB are from archaea. In five organisms, one archaeon (*Nanoarchaeum equitans*) and four bacteria (*Borrelia burgdorferi*, *Chlamydia trachomatis*, *Mesoplasma florum*, and *Onion yellows phytoplasma*), our search method did not retrieve any putative copper protein. However, all these examples are host-associated organisms, so we can speculate that they exploit the copper proteome of their host. The organism with the largest percentage of copper-binding proteins is *Nitrosomonas europaea*, an autotrophic nitrifying bacterium which, with respect to other organisms, has a large number of copper-dependent oxidoreductase enzymes encoded in a relatively small proteome. Halobacterium sp. is the most rich in copper proteins among archeal organisms; in particular, it has the largest archaeal repertoire of proteins belonging to the plastocyanin/azurin family. Among the eukaryota investigated here, the largest proteome fraction of copper proteins is observed in *Arabidopsis thaliana* and *Saccharomyces cerevisiae.* The *Arabidopsis* proteome has a larger content of Cu-oxidoreductase enzymes with respect to the other eukaryotes. The number of proteins identified as a function of the proteome size is plotted in Figure 2B. It can be seen that the number of putative copper-binding proteins does not correlate well with the size of the proteome in all three domains of life. Some proportionality is observed for archaea and eukaryota, albeit with significant deviations, whereas there is little, if any, correlation in the case of bacteria. This is at variance with other metalloproteomes, as highlighted in previous studies. For instance, the content of zinc proteins has been shown to be dependent on the proteome size.[19]

The putative copper-binding proteins identified have been grouped in 56 ensembles (Table S7), as described in the Methods section. 49 proteins could not be assigned to any ensemble (Table S8). Each ensemble may contain proteins from all three domains of life, from two domains, or from one domain only. The share of proteins belonging to ensembles spanning all three domains of life is relatively high. In particular, 71% and 59% of the putative copper proteins retrieved in bacterial and archaeal organisms, respectively, have homologues in all three domains of life (Figure 3). This suggests that, at least for bacterial organisms, the largest part of their copper-binding protein repertoire is probably evolutionarily ancient and may have been vertically inherited from the so-called last common ancestor of all organisms. In the case of eukaryotic organisms, a non-negligible fraction of copper proteins (34%) has homologues in the three domains of life; however, the largest percentage (57%) of the copper proteome is specific for this superkingdom. Notably, ca. 30% of archaeal copper proteins belongs to two families that contain also bacterial homologues (Figure 3). One family contains 5 proteins, whereas the other one contains 35 proteins. The ratio of bacterial/archaeal members in the largest of these families, which contains proteins with a NosD domain, is 1:10 (in *Methanosarcina acetivorans* the NosD domain is found in 26 proteins). This, together with the significantly higher number of bacterial vs archaeal copper proteins detected, causes the fraction of

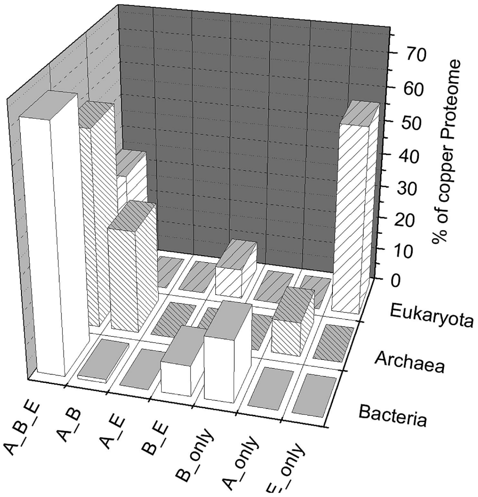

Figure 3. Fraction (with respect to the entire copper proteome of the respective domain of life) of putative copper-binding proteins in each domain of life that belongs to ensembles whose members span all three, two, or only one domain (A, archaea; B, bacteria; E, eukaryota).

bacterial proteins with archaeal homologues to be only 2% (Figure 3). Inspection of Figure 4 indicates that eukaryotic organisms had both the capacity to evolve new copper domains, which are then specific to their superkingdom (e.g., Ctr), and the ability to develop new roles for the ancient copper domains common to all the superkingdoms. The latter instance holds, for instance, for the CCS copper chaperone, which is a eukaryotic protein involved in copper insertion in Cu/Zn superoxide dismutase and which is formed by a fusion of two widespread copper domains (HMA and Sod_Cu). Furthermore, eukaryotes make a larger use of domains of the type of multicopper oxidases (Cu-oxidase, Cu-oxidase_2, Cu-oxidase_3 in Figure 4) and of heavy metal transport and detoxification proteins (HMA) than prokaryotes. The plant *Arabidopsis thaliana* is particularly rich in these domains (about 50 copies), also with respect to other eukaryotic organisms which have a content ranging between 1 and 12 copies. It must be noted that, at least in the case of the retrieved proteins with the HMA domain, it is possible that some percentage of this ensemble does not bind copper in vivo, because the HMA domain can bind different metals simply adding some ligands to those used to bind copper (e.g., with a DCXXC pattern rather than CXXC). The above observations are in general agreement with the larger number of putative copper proteins retrieved in eukaryota (an average of 76) with respect to that found in archaea and bacteria (an average of 10). This leads us to conclude that eukaryotic organisms have both an ancient copper protein repertoire and a relatively recent copper protein portfolio probably evolved to answer the demand of controlling the concentration of free copper and its transport in more and more complex systems such as multicellular organisms. In the latter organisms, it is well-known that some copper-binding proteins play key roles in the formation of the extracellular matrix. For example, the activity of lysyl oxidases is required for the normal maturation and cross-linking of collagen and elastin.[42] Deficient levels of lysyl oxidase-dependent cross-links in bone collagen have been associated with osteoporosis and weak bones.[43,44] Ceruloplasmin, a copper-dependent ferroxidase, was recently reported to have a specific aggregative action on young neurons in vitro, having a potential role in the development of the nervous system organization.[45]

4. Functional Implications. It is intriguing to analyze, at least at a relatively general level, which is the range of functions carried out by the putative copper proteins identified. Not unexpectedly, the analysis of the GO terms associated to the copper proteins retrieved indicates that the most common functional roles for copper-binding proteins in organisms are essentially two, i.e., the homeostasis of copper (between 38% and 45%) and the catalysis of redox reactions or electron transfer (between 37% and 49%). 1% of eukaryotic copper proteins is involved in the transport of oxygen. These proteins belong to the family of hemocyanins, which are copper-containing proteins found in the hemolymph of many invertebrates. Due to its redox properties, copper is one of the essential trace elements, exploited by organisms to perform oxidation–reduction reactions and electron transfer. However, free copper is highly toxic for cells,[46] due to its ability to promote reactions such as generation of reactive oxygen species via redox cycling. These potentially lethal properties caused all the organisms to evolve sophisticated copper homeostasis mechanisms that regulate uptake, distribution, sequestration, and export of copper, limiting the free copper concentration.[47,48]

Many eukaryotic-specific copper-binding ensembles (9 out of 27) have unknown function. This explains why the percentage of copper proteins with unknown function is much higher in eukaryotes than in prokaryotes. In this regard, it is important to note that the "CHCH" domain, which is a eukaryotic domain with unknown function, was included in this search as a copper-binding domain, although it is not annotated as such and it does not have any copper-containing structure. The inclusion of this domain among the copper-binding domains is based on the observation that it is found in the copper-binding protein Cox19.[24] The detailed analysis of the subcellular localization of proteins can also be functionally informative (Tables S4–S6, Supporting Information). In this respect, however, the quality of the predictions that can be obtained from currently available software tools is limiting (see also Methods). Thus, a detailed discussion is worthy only for human proteins, for which a database of experimentally verified localizations is available.[37] It is observed that cytosolic copper proteins are essentially all devoted to copper transport, storage, or sensing, with the relevant exception of superoxide dismutase 1 (SOD1). Consistently with this, and in qualitative agreement also with the reduction potential of the cytosol that stabilizes copper(I), the large majority of "cytosolic CBPs" contains only cysteines. CBPs without cysteines are detected only in SOD and its copper chaperone.[49] A significant share of copper proteins localizes to the mitochondrion: this includes subunits I and II of cytochrome *c* oxidase and several proteins involved in the assembly of their copper sites. Other membrane-bound proteins include both copper transporters, whose CBPs contain mainly cysteines or methionines and copper-containing enzymes, such as tyrosinase. Finally, some human extracellular copper proteins have been mentioned in the preceding section.

Out of the 56 copper protein ensembles identified in this work, eight span all three domains of life. When a family of proteins has many representatives in all domains of life, then it is commonly assumed that it was present in the last common ancestor.[50] On this basis and assuming that our copper ensembles represent families of orthologues, the eight copper ensembles present in all domains of life were analyzed looking for potential functional associations among these ensembles, as described in the Methods section. To this aim, a COG

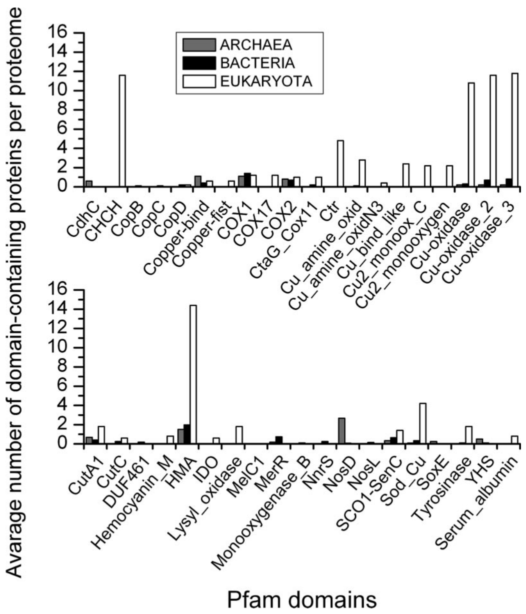

Figure 4. Fraction of putative copper-binding proteins containing a given Pfam domain in the total copper proteome of archaeal (gray), bacterial (black), and eukaryotic (white) organisms.

identification number was associated to each ensemble (see Figure 5). According to the STRING database, six of these ensembles are interconnected by high-confidence functional associations (score ≥ 0.7), as shown in Figure 5. All the links between these clusters included gene neighborhood in at least five organisms. Note that Figure 5 does not depict a pathway in which, for example, the copper ions are subsequently transferred from one protein to the next, but rather represents a combination of different pieces of evidence that there are relationships of a functional nature between the protein ensembles shown. In fact, linkages between pairs of these proteins as well as parts of this network have been extensively investigated by several research groups worldwide.[51,52] However, these interactions are not all simultaneously present, depending on each organism's protein content and cellular structure. In bacteria, proteins of the so-called Sco/SenC family are involved in the assembly of Cu_A-containing cytochrome *c* oxidases, which catalyze dioxygen reduction, but are also likely to have a role in copper incorporation in multicopper oxidases[53] (Figure 5). Copper-transporting ATPases are also functionally involved in the incorporation of the copper ion into multicopper oxidases, e.g., in eukaryotes through the well-studied pathway for copper transport into the trans-Golgi network[54] where the metal is incorporated in copper oxidases but also in bacteria where there are a few instances of close proximity in the genome of ATPase and oxidase genes. Altogether, the ensembles of proteins shown in Figure 5 and their functional interactions constitute a major determinant of cellular copper homeostasis, e.g., in Gram-negative bacteria, by establishing the level of copper usage in the periplasmic space and by controlling the quantity of copper present in the cytoplasm.

Conclusions

In the present work, we predicted the copper proteome of organisms selected from the three domains of life. We found that the size of the copper proteome is generally less than 1% of the total proteome of an organism, for both eukaryotes and prokaryotes. Grouping of the copper proteins identified into ensembles of homologues revealed that a relatively small set of them constitutes the majority of the copper proteome in prokaryotes and are also common to eukaryotes. Functional

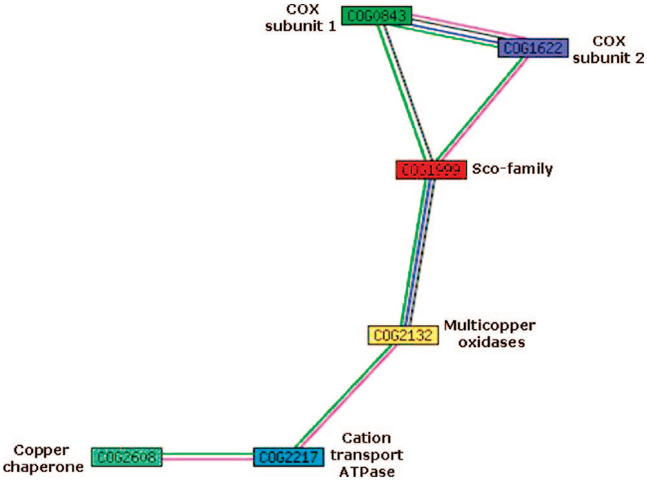

Figure 5. Potential functional associations predicted for six out of eight copper ensembles spanning all domains of life. The analysis was performed by associating a COG identification number to each of these ensembles and searching the String database in the COG mode. Evidence for a functional association between copper chaperones (COG2608) and cation transport ATPases (COG2217) results from gene neighborhood analysis (green line) and experimental data (pink line). Evidence for a functional association between cation transport ATPases (COG2217) and multicopper oxidases (COG2132) results from gene neighborhood analysis (green line) and experimental data (pink line). Evidence for a functional association between multicopper oxidases (COG2132) and Sco proteins (COG1999) results from gene neighborhood analysis (green line) and statistical data on their co-occurrence in genomes (blue line). Evidence for a functional association between Sco proteins (COG1999) and subunits I and II of cytochome *c* oxidase results from gene neighborhood analysis (green lines), experimental data (pink line), and coexpression data (brown line). Subunits I and II of cytochrome *c* oxidase are linked by several pieces of evidence, including structural data.

predictions further indicated that these proteins are likely to be part of a network which may thus represent an ancient core that is crucial for copper homeostasis. It appears that the differentiation of prokaryotic organisms affected only slightly this ancestral copper proteome. On the other hand, eukaryotes have expanded their ancestral repertoires of copper proteins, by both inventing new copper domains and reusing old domains for new functions. The main driving force for this evolution is most likely the necessity to tightly regulate the homeostasis of Cu(I), which is essential for life but is also toxic to cells in the free form, in the more complex eukaryotic organisms.

Abbreviations: MBP, metal-binding pattern; EC, enzyme classification; CBP, copper-binding pattern.

Acknowledgment. This work was supported by the European Commission (Contract no. 031220, "Spine2-complexes") and by MIUR (Contract no. 2005039878, "PRIN 2005").

Supporting Information Available: Figures S1–S4 and Tables S1–S8. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Linder, M. C. *Biochemistry of Copper*; Plenum Press: NY, 1991; pp 1–13.

(2) Vulpe, C. D.; Packman, S. *Ann. Rev. Nutrition* **1995**, *15*, 293–322.

(3) O'Halloran, T. V.; Culotta, V. C. *J. Biol. Chem.* **2000**, *275*, 25057–25060.

(4) Harrison, M. D.; Jones, C. E.; Solioz, M.; Dameron, C. T. *Trends Biochem. Sci.* **2000**, *25*, 29–32.

(5) Puig, S.; Thiele, D. J. *Curr. Opin. Chem. Biol.* **2002**, *6*, 171–180.

(6) Huffman, D. L.; O'Halloran, T. V. *J. Biol. Chem.* **2000**, *275*, 18611–18614.

(7) Rae, T.; Schmidt, P. J.; Pufahl, R. A.; Culotta, V. C.; O'Halloran, T. V. *Science* **1999**, *284*, 805–808.

(8) Klomp, L. W.; Lin, S. J.; Yuan, D.; Klausner, R. D.; Culotta, V. C.; Gitlin, J. D. *J. Biol. Chem.* **1997**, *272*, 9221–9226.

(9) Pufahl, R. A.; Singer, C. P.; Peariso, K. L.; Lin, S.-J.; Schmidt, P. J.; Fahrni, C. J.; Cizewski Culotta, V.; Penner-Hahn, J. E.; O'Halloran, T. V. *Science* **1997**, *278*, 853–856.

(10) Petris, M. J.; Mercer, J. F.; Culvenor, J. G.; Lockhart, P.; Camakaris, J. *EMBO J.* **1996**, *15*, 6084–6095.

(11) Arnesano, F.; Banci, L.; Bertini, I.; Cantini, F.; Ciofi-Baffoni, S.; Huffman, D. L.; O'Halloran, T. V. *J. Biol. Chem.* **2001**, *276*, 41365–41376.

(12) Banci, L.; Bertini, I.; Cantini, F.; Chasapis, C.; Hadjiliadis, N.; Rosato, A. *J. Biol. Chem.* **2005**, *280*, 38259–38263.

(13) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(14) Andreini, C.; Bertini, I.; Rosato, A. *Bioinformatics* **2004**, *20*, 1373–1380.

(15) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535–542.

(16) Zhang, Z.; Schaffer, A. A.; Miller, W.; Madden, T. L.; Lipman, D. J.; Koonin, E. V.; Altschul, S. F. *Nucleic Acids Res.* **1998**, *26*, 3986–3990.

(17) Andreini, C.; Banci, L.; Bertini, I.; Rosato, A. *J. Proteome Res.* **2006**, *5*, 196–201.

(18) Andreini, C.; Banci, L.; Bertini, I.; Elmi, S.; Rosato, A. *Proteins* **2007**, *67*, 317–324.

(19) Andreini, C.; Banci, L.; Bertini, I.; Rosato, A. *J. Proteome Res.* **2006**, *5*, 3173–3178.

(20) Passerini, A.; Andreini, C.; Menchetti, S.; Rosato, A.; Frasconi, P. *BMC Bioinf.* **2007**, *8*, 39.

(21) Lin, H. H.; Han, L. Y.; Zhang, H. L.; Zheng, C. J.; Xie, B.; Cao, Z. W.; Chen, Y. Z. *BMC Bioinf.* **2006**, *7* (5), S13.

(22) Al-Karadaghi, S.; Cedergren-Zeppezauer, E. S.; Dauter, Z.; Wilson, K. S. *Acta Crystallogr, Sect. D.: Biol. Crystallogr.* **2007**, *51*, 805–813.

(23) Bateman, A.; Coin, L.; Durbin, R.; Finn, R. D.; Hollich, V.; Griffiths-Jones, S.; Khanna, A.; Marshall, M.; Moxon, S.; Sonnhammer, E. L.; Studholme, D. J.; Yeats, C.; Eddy, S. R. *Nucleic Acids Res.* **2004**, *32 Database issue*, D138–D141.

(24) Rigby, K.; Zhang, L.; Cobine, P. A.; George, G. N.; Winge, D. R. *J. Biol. Chem.* **2007**, *282*, 10233–10242.

(25) Mothes, E.; Faller, P. *Biochemistry* **2007**, *46*, 2267–2274.

(26) Keil, H. L.; Nelson, V. E. *J. Biol. Chem.* **1931**, *93*, 49–57.

(27) Bonaventura, C.; Godette, G.; Tesh, S.; Holm, D. E.; Bonaventura, J.; Crumbliss, A. L.; Pearce, L. L.; Peterson, J. *J. Biol. Chem.* **1999**, *274*, 5499–5507.

(28) Eddy, S. R. *Bioinformatics* **1998**, *14*, 755–763.

(29) McGinnis, S.; Madden, T. L. *Nucleic Acids Res.* **2004**, *32*, W20–W25.

(30) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. *Nat. Genet.* **2000**, *25*, 25–29.

(31) Tatusov, R. L.; Koonin, E. V.; Lipman, D. J. *Science* **1997**, *278*, 631–637.

(32) Snel, B.; Lehmann, G.; Bork, P.; Huynen, M. A. *Nucleic Acids Res.* **2000**, *28*, 3442–3444.

(33) von Mering, C.; Huynen, M.; Jaeggi, D.; Schmidt, S.; Bork, P.; Snel, B. *Nucleic Acids Res.* **2003**, *31*, 258–261.

(34) Gardy, J. L.; Laird, M. R.; Chen, F.; Rey, S.; Walsh, C. J.; Ester, M.; Brinkman, F. S. *Bioinformatics.* **2005**, *21*, 617–623.

(35) Rey, S.; Acab, M.; Gardy, J. L.; Laird, M. R.; deFays, K.; Lambert, C.; Brinkman, F. S. *Nucleic Acids Res.* **2005**, *33*, D164–D168.

(36) Horton, P.; Park, K. J.; Obayashi, T.; Fujita, N.; Harada, H.; Adams-Collier, C. J.; Nakai, K. *Nucleic Acids Res.* **2007**, *35*, 585–587.

(37) Fink, J. L.; Aturaliya, R. N.; Davis, M. J.; Zhang, F.; Hanson, K.; Teasdale, M. S.; Kai, C.; Kawai, J.; Carninci, P.; Hayashizaki, Y.; Teasdale, R. D. *Nucleic Acids Res.* **2006**, *34*, 213–217.

(38) *Handbook of Metalloproteins*; Wiley: Chichester (UK), 2001; pp 1–1248.

(39) Banci, L.; Bertini, I.; Calderone, V.; Ciofi-Baffoni, S.; Mangani, S.; Martinelli, M.; Palumaa, P.; Wang, S. *Proc. Natl. Acad. Sci. USA* **2006**, *103* (31), 8595–8600.

(40) Imriskova-Sosova, I.; Andrews, D.; Yam, K.; Davidson, D.; Yachnin, Y.; Hill, B. C. *Biochemistry* **2006**, *44*, 16949–16956.

(41) (a) Horng, Y. C.; Leary, S. C.; Cobine, P. A.; Young, F. B.; George, G. N.; Shoubridge, E. A.; Winge, D. R. *J. Biol. Chem.* **2005**, *280*, 34113–34122. (b) Banci, L.; Bertini, I.; Cantini, F.; Felli, I. C.; Gonnelli, L.; Hadjiliadis, N.; Pierattelli, R.; Rosato, A.; Voulgaris, P. *Nat. Chem. Biol.* **2006**, *2* (7), 367–368.

(42) Guo, Y.; Pischon, N.; Palamakumbura, A. H.; Trackman, P. C. *Am. J. Physiol. Cell Physiol.* **2007**, *292*, C2095–C2102.

(43) Bailey, A. J.; Wotton, S. F.; Sims, T. J.; Thompson, P. W. *Connect. Tissue Res.* **1993**, *29*, 119–132.

(44) Oxlund, H.; Mosekilde, L.; Ortoft, G. *Bone* **1996**, *19*, 479–484.

(45) Maltais, D.; Desroches, D.; Aouffen, M.; Mateescu, M. A.; Wang, R.; Paquin, J. *Neuroscience* **2003**, *121*, 73–82.

(46) Beswick, P. H.; Hall, G. H.; Hook, A. J.; Little, K.; McBrien, D. C.; Lott, K. A. *Chem. Biol. Interact.* **1976**, *14*, 347–356.

(47) Cavet, J. S.; Borrelly, G. P.; Robinson, N. J. *FEMS Microbiol. Rev.* **2003**, *27*, 165–181.

(48) Tottey, S.; Harvie, D. R.; Robinson, N. J. *Acc. Chem. Res.* **2005**, *38*, 775–783.

(49) Lamb, A. L.; Torres, A. S.; O'Halloran, T. V.; Rosenzweig, A. C. *Nat. Struct. Biol.* **2001**, *8*, 751–755.

(50) Doolittle, W. F. *Curr. Opin. Struct. Biol.* **2000**, *10*, 355–358.

(51) Rosenzweig, A. C.; O'Halloran, T. V. *Curr. Opin. Chem. Biol.* **2000**, *4*, 140–147.

(52) Luk, E.; Jensen, L. T.; Culotta, V. C. *J. Biol. Inorg. Chem.* **2003**, *8*, 803–809.

(53) Banci, L.; Bertini, I.; Cavallaro, G.; Rosato, A. *J. Proteome Res.* **2007**, *6*, 1568–1579.

(54) Bertini, I.; Rosato, A. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 3601–3604.

PR070480U