ACCOUNTS
of chemical research

Metalloproteomes: A Bioinformatic Approach

CLAUDIA ANDREINI,[†] 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 ON JANUARY 12, 2009

CONSPECTUS

Genome-wide studies are providing researchers with a potentially complete list of the molecular components present in living systems. It is now evident that several metal ions are essential to life and that metalloproteins, that is, proteins that require a metal ion to perform their physiological function, are widespread in all organisms. However, there is currently a lack of well-established experimental methods aimed at analyzing the complete set of metalloproteins encoded by an organism (the metalloproteome). This information is essential for a comprehensive understanding of the whole of the processes occurring in living systems. Predictive tools must thus be applied to define metalloproteomes.

In this Account, we discuss the current progress in the development of bioinformatics methods for the prediction, based solely on protein sequences, of metalloproteins. With these methods, it is possible to scan entire proteomes for metalloproteins, such as zinc proteins or copper proteins, which are identified by the presence of specific metal-binding sites, metal-binding domains, or both. The predicted metalloproteins can be then analyzed to obtain information on their function and evolution. For example, the comparative analysis of the content and usage of different metalloproteins across living organisms can be used to obtain hints on the evolution of metalloproteomes.

As case studies, we predicted the content of zinc, nonheme iron, and copper-proteins in a representative set of organisms taken from the three domains of life. The zinc proteome represents about 9% of the entire proteome in eukaryotes, but it ranges from 5% to 6% in prokaryotes, therefore indicating a substantial increase of the number of zinc proteins in higher organisms. In contrast, the number of nonheme iron proteins is relatively constant in eukaryotes and prokaryotes, and therefore their relative share diminishes in passing from archaea (about 7%), to bacteria (about 4%), to eukaryotes (about 1%). Copper proteins represent less than 1% of the proteomes in all the organisms studied.

We also discuss the limits of these methods, the approaches used to overcome some of these limits to improve our predictions, and possible future developments in the field of bioinformatics-based investigation of metalloproteins. As a longstanding goal of the biological sciences, the understanding of life at the systems level, or systems biology, is experiencing a rekindling of interest; ready access to complete information on metalloproteomes is crucial to correctly represent the role of metal ions in living organisms.

Introduction

Life on Earth developed in equilibrium with the hydrosphere and the lithosphere, taking from these all the elements necessary for performing essential functions. As a consequence, a number of metal ions have been selected during evolution to take part in many crucial biological processes and are thus essential for living organisms.[1] In particular, many proteins require metal ions to carry out their physiological functions.[2] We propose to refer to them as metalloproteins. Up to now, there was discrimination between metal—protein complexes and metalloproteins, depending on the affinity constant. In the present definition, the designation of metalloproteins is used for proteins that require a metal ion or metal-con-

Published on the Web 08/21/2009 www.pubs.acs.org/acr
10.1021/ar900015x CCC: $71.50

taining cofactor for functional or structural reasons. An analysis of enzyme mechanisms, restricted to enzymes with known structure, has shown that about 40% of enzyme-catalyzed reactions involve metal ions.[3]

Genome sequencing and postgenomic projects have been providing researchers with a potentially complete list of the components that are present in organisms and of the relationships between them. Such an unprecedented wealth of information has led to the renaissance of a long-standing goal of biological sciences, namely, the understanding of living organisms at the systems level, or systems biology.[4] Systems biology aims at describing the behavior of biological systems based on their molecular constituents, and thus it cannot be abstracted from the investigation of metals and metalloproteins because they are cellular components of crucial importance. Systems biology approaches require the combination of large-scale studies to catalogue genome-wide data sets to obtain as detailed as possible knowledge on the molecules and their interactions. The investigation of metalloproteins in this framework, therefore, implies the definition of all the metalloproteins encoded by an organism (which constitute its metalloproteome) in conjunction with their functional characterization. Bioinformatics methods can give valuable support to experimental methods in both of these efforts and are especially important to obtain insights into metalloproteomes (metal by metal), given the fact that high-throughput experimental technologies for their characterization are not yet routinely available.[5]

In this Account, we discuss the development of bioinformatics methods focused on the prediction of metalloproteins, metal by metal, and we show that they make it possible to deduce metalloproteomes of living organisms. The analysis of the content of zinc-, copper-, and iron-binding proteins in representative organisms has provided hints to understand several properties of these metalloproteins, as well as the usage of these metal ions within each domain of life and the functions of the corresponding metalloproteins. The same analysis could be extended also to the study of other metal ions. We also describe current limitations of these methods and possible future work to improve them. Despite their limitations, bioinformatic approaches represent an important contribution to the overall comprehension of how metal ions are framed as essential factors in living systems.

Methods for Metalloproteome Prediction

Metalloproteins, as defined in the Introduction, are identified through biochemical studies that probe the dependence of the function of the proteins of interest on the presence of metal ions. This is typically done *in vitro* on purified native or recombinant samples. To obtain details on the properties of the metal site(s), one needs high-quality samples on which a variety of analytical and spectroscopic tools are applied or which are used for a complete protein structure determination. Therefore, the identification and characterization of a metalloprotein is the result of the successful completion of a relatively long series of complex tasks, which often can be quite challenging. At present, the complexity of and the resource demands associated with the needed experimental work make it unfeasible to perform a complete identification of metalloproteins at the level of entire metalloproteomes. Therefore, there is a need to develop methods for the prediction of metal-binding sites on the basis of the protein sequence only. To this aim, we proposed to identify metalloproteins by searching for known metal-binding domains in their sequences. Lists of metal-binding domains can be extracted from libraries such as Pfam.[6] The detailed procedure is described in the next paragraph and depicted in Figure 1.[7]

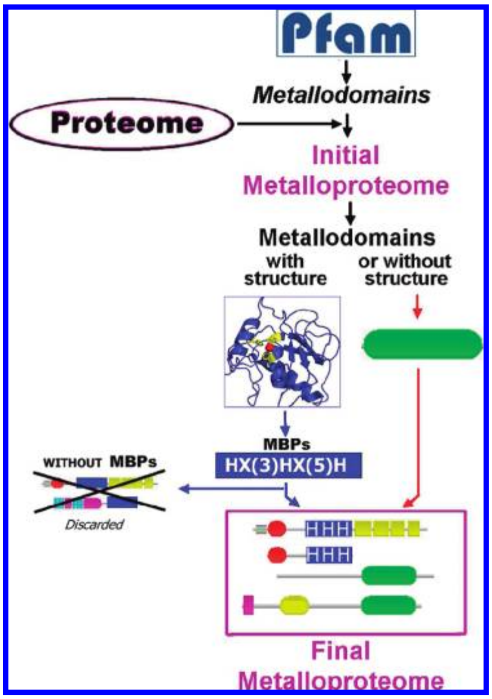

FIGURE 1. Schematic representation of the approach for the detection of metalloproteins in complete proteomes.[7,11,12]

In domain-based searches, we look for the occurrence of known metal-binding domains in the protein sequences. A domain is a structurally and functionally defined protein region that can be characterized by a multiple sequence alignment. This information is typically condensed in a so-called sequence profile, that is, a table reporting for each position in the protein sequence the likelihood of the occurrence of any of the natural amino acids as well as of sequence gaps. Using the sequence profiles available in the Pfam library, every proteome in GenBank is analyzed with the search tool HMMER.[8] The relevant profiles in Pfam, that is, those corresponding to metal-binding domains, can be selected by querying the library for those whose annotation contains the metal name or symbol and then checking the primary literature to discard domains erroneously retrieved. In this way, for each metal, we find an ensemble of proteins that can be called the predicted metalloproteome. This procedure has been already reported.[7] This predicted metalloproteome however suffers from two limitations. First, a number of proteins may have lost the metal-coordinating ligands (i.e., they may lack the amino acids of the profile that are responsible for binding the metal ion) during evolutionary pressure; therefore they cannot be metalloproteins unless a new, uncharacterized metal-binding site has formed concomitantly. From our experience, this error may be as large as 30% of the ensemble. Second, there are a number of domains that are not yet recognized as metal-binding because the annotation of Pfam is largely done manually. This introduces an error smaller than the first point; this bias can be estimated as less than 5% of the whole predicted metalloproteome. We can substantially reduce the errors due to the aforementioned reasons by introducing the concept of a metal-binding pattern (MBP). The latter may be defined as the sequence motif found in the structure of the protein from the PDB responsible for binding the metal cofactor.[9] It can be represented by specifying the metal ligands and their spacing in the sequence (e.g., HX(3)HX(5)H, three histidine ligands separated by any three and five residues in the sequence, respectively). For metalloproteins of known three-dimensional structure, available from the PDB, the program HMMER is used to recognize the Pfam metal-binding profiles not already annotated as such, recovering the missing 5% mentioned before. In the future, with the annotation of Pfam becoming more complete, the weight of this correction will reduce. Instead, in order to reduce the overestimate due to the inclusion in the predicted metalloproteome of proteins corresponding to the profile but lacking the capability of binding metal ions, we divide the metalloproteome into subgroups, according to whether the metal-binding profile from which each metalloprotein was retrieved has at least one representative with known three-dimensional structure. The relative size of the two subgroups is typically about 70:30 (with structure/without structure). At this point, we filter off those proteins that do not have the MBP. The filter is applied by imposing that the predicted metalloprotein contains all the ligands of the MBP with spacing in sequence that it is maintained within ±20% (or ±1 amino acid for short spacing). The proteins filtered off are about 30%,[7] leading to an improvement of the average precision of the methodology from about 50% to about 85%. Precision is defined as the ratio between the number of the proteins correctly predicted as metalloproteins over the total number of predicted metalloproteins, that is, is the fraction of correct predictions. We have not conceived any procedure to perform an analogous filter for the subgroup for which no three-dimensional structure is known. However, considering that the relative amount of these profiles is 30% and that the overestimate is 30%, the final output is overestimated by 9%. With time, the number of profiles without structure is going to decrease, and consequently the overestimate is expected to decrease.

Even though the procedure described in the previous paragraph tends to slightly overestimate the number of metalloproteins, it may miss some of them when the profile is particularly short in sequence (less than 40−50 amino acids). This specific point may be addressed through an alternative procedure, reported in refs 10−12, based on the PHI-BLAST program. This procedure is however globally less sensitive than the previously described protocol. Typically, an additional 5% of metalloproteins can be identified. Finally, completely unprecedented, uncharacterized metalloproteins can be predicted by using support vector machines (SVMs),[13−16] which however are generally more error-prone and less comprehensive.

A Case Study: Comparative Analysis of Zinc Proteomes

A search for zinc proteomes in 57 representative organisms from the three domains of life (40 bacteria, 12 archaea, and 5 eukaryotes) is available in the literature,[7] as obtained through Pfam, the filter for MBPs and the use of the PHI-BLAST program, as indicated in the previous section. It is shown that zinc proteins are widespread in living organisms. Within each domain of life, there exists a good correlation between the zinc-protein content and the proteome size of the organism (Figure S1, Supporting Information[7]). Prokaryotic organisms, on average, have a lower fraction of zinc proteins (6.0% ± 0.2% of the entire proteome in archaea and 4.9% ± 0.1% in bacteria) than eukaryotic organisms. The zinc proteome in fact constitutes, on average, 8.8% ± 0.4% of the eukaryotic proteome (about 10% in humans), thus representing a much more important fraction. Approximately, two-thirds of the prokaryotic zinc proteins have homologues in eukaryotes, while the remaining third comprises zinc proteins encoded only in prokaryotic organisms. On average, three-quarters of the eukaryotic zinc proteomes comprise proteins encoded

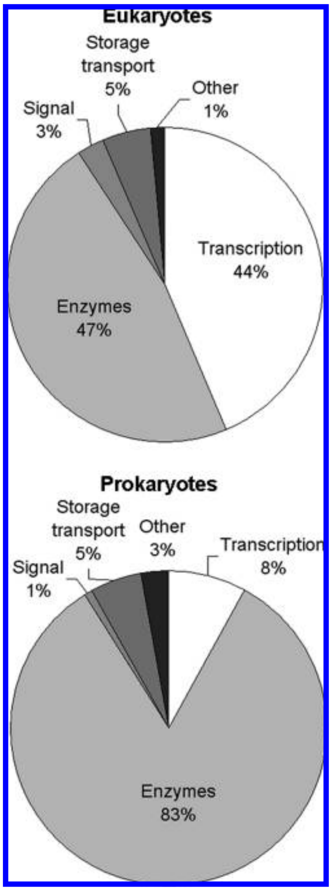

FIGURE 2. Distribution of the functions of zinc proteins in eukaryotes and prokaryotes. The graph includes only the retrieved proteins with known function, which represent about 90% of the total.[7]

only in eukaryotes, suggesting that they are relatively more recent in evolution.[7]

As shown in Figure 2,[7] there is also a functional diversification of the eukaryotic and prokaryotic zinc proteomes. Prokaryotes use zinc proteins to perform enzymatic catalysis, whereas in eukaryotes the zinc proteome is almost equally involved in performing catalysis and in regulating DNA transcription. This broad difference in function has a correspondence with the organization of the zinc-binding patterns. Indeed, the patterns containing four ligands are associated with structural sites, that is, zinc contributes to the stability of the protein structure, whereas zinc-binding patterns containing three protein ligands are associated with catalytic sites, that is, zinc actively participates in the reaction mechanism of the enzyme.[17] In the latter case, the metal ion most often completes its first coordination sphere by binding a water molecule, which participates in the reaction mechanism, or a substrate molecule. The identity of the amino acids in the pattern is also quite different among structural and catalytic sites. In humans, 97% of the proteins having a structural zinc site contain at least one cysteine ligand, with 40% having four cysteine ligands in their MBP.[10] On the other hand, nearly one-third of the human proteins with a three-ligand MBP have a pattern with three histidines. Together, four- and three-ligand patterns account for ca. 96% of all human zinc proteins. It is useful to compare these results with the data maintained in other databases. A suitable example is that of the MEROPS database,[18] which contains information on peptidases, including metallopeptidases. For zinc peptidases, the agreement is large: only a few proteins of the latter database are missing in our outputs because they lack a MBP, as defined by us.

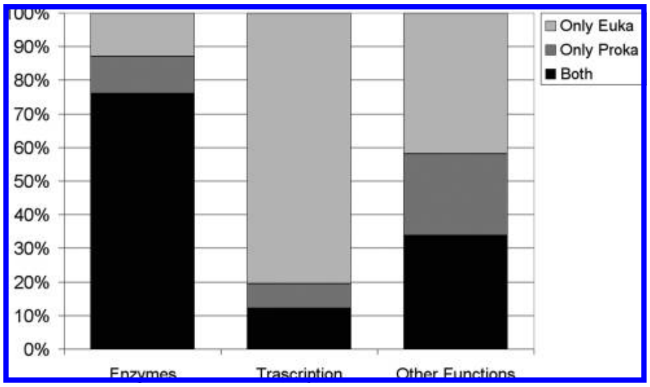

FIGURE 3. Fraction of zinc proteins having homologues in both prokaryotes and eukaryotes or being specific to a superdomain. The value for enzymes, transcription factors, and proteins performing other kinds of functions are shown separately.[7]

Figure 3[7] shows that most zinc-dependent enzymes have homologues in both prokaryotes and eukaryotes. This suggests that zinc has been exploited in the catalytic site of enzymes before the differentiation of the organisms in the three domains of life. On the other hand, zinc-binding transcription factors are almost exclusively a prerogative of eukaryotes (Figures 2 and 3). These transcription factors commonly contain zinc-finger domains, which are much rarer in bacteria and archaea.[19] So, whereas zinc enzymes seem to come from a more ancient zinc proteome, zinc-binding transcription factors seem to have appeared to meet the need of higher organisms to regulate processes like cell compartmentalization and, for multicellular organisms, cell differentiation. This hypothesis is supported also by the results of the analysis of the MBPs. Transcription factors bind zinc in very similar metal sites, most often composed by cysteine and histidine and organized in the same 3D structure.[20] Instead, enzymes use a larger variety of zinc-binding sites. The conservation of zinc-finger binding sites could be associated with their more recent origin, whereas the differentiation of the catalytic zinc-binding sites could be the result of evolutionary processes that resulted in the development of different enzymatic reactions targeting different physiological substrates. Indeed, prokaryotic zinc proteins without homologues in eukaryotes are most often specific to only a few bacterial classes and may be the result of environmental adaptation. In this regard, it is worth mentioning that the zinc proteome of the 11 hyperthermophilic organisms studied constitutes the 7.0% ± 1.1% of the entire proteome, compared with 6.0% ± 1.0% of 5 thermophilic organisms, 5.3% ± 1.0% of 34 mesophilic organisms, and 4.5% ± 0.2% of 2 psychrophilic organisms.[7] This effect may be due to an increased use of zinc to enhance the structural stability of proteins by organisms living at higher temperatures.

Application to Other Transition Metal Ions

The methods described for zinc proteomes can be applied to the study of all metalloproteins that use transition metal ions, including less frequent metals like molybdenum and tungsten.

Proteome-level analyses of the occurrence of nonheme iron proteins have shown that, at variance with what was observed for zinc, there is no expansion of the nonheme iron proteome in eukaryotes with respect to prokaryotes.[11] Nonheme iron proteins constitute, on average, 7.1% ± 2.1% of archaeal proteomes, 3.9% ± 1.6% of bacterial proteomes, and only 1.1% ± 0.4% of eukaryotic proteomes. The majority of these proteins have homologues in all three domains of life (about 90% of the total) suggesting that the existing organisms share the bulk of nonheme iron proteins, which possibly appeared early in the course of evolution. In this regard, the different percentages can be explained by the different proteome sizes, given that the same bulk of proteins is "diluted" in larger eukaryotic proteomes. The large majority of nonheme iron sites are found in proteins involved in electron transfer or in enzymes performing oxidoreductase activity (Figure 4, top[11]). Iron is the most used metal ion in redox catalysis, followed by copper and molybdenum.[3] All these metal ions have at least two oxidation states that can be sufficiently stabilized in proteins (+2, +3, and +4 in the case of iron), so that metal ions can cycle between them during catalysis. Iron is the metal ion with the largest variety of sites in proteins, including several kinds of iron−sulfur clusters and heme cofactors. This may be due to the necessity to use different chemical environments to modulate the reduction potential of iron and thus its reactivity. Iron−sulfur clusters are the cofactor of about 40% of the nonheme iron proteins retrieved, and their binding patterns are most often composed of cysteine residues. It is worth noting that cysteine is conversely an uncommon ligand for all the other nonheme iron sites, where histidine is the most widespread ligand.

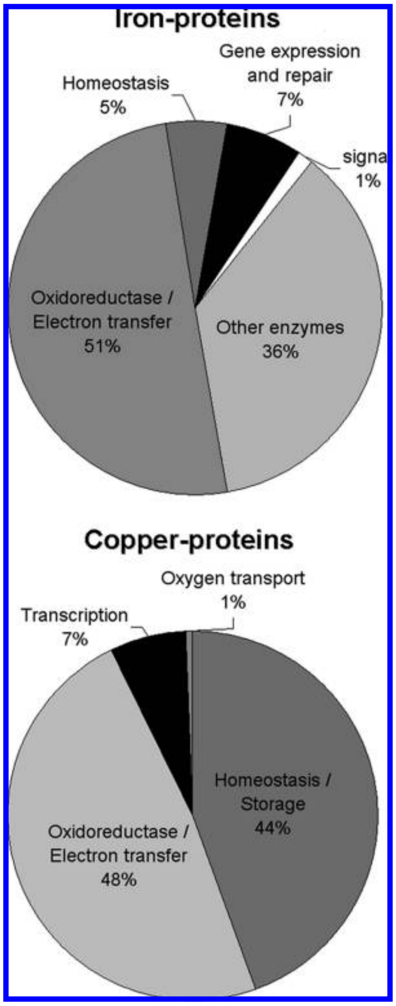

FIGURE 4. Distribution of the function of nonheme iron[11] (top) and copper proteins[12] (bottom). The graph includes the retrieved proteins with known function, which represent about 85% and 90% of the total, respectively.

Copper is the second most common ion in redox catalysis[3] and cycles between the +1 and +2 states. Another common role for copper proteins is in copper homeostasis (Figure 4, bottom).[12] Because of the potential toxicity of this ion for the cell, Nature has evolved complex systems to control the intracellular concentration of copper.[21−23] In these proteins, copper is most often bound in its +1 state by cysteine residues. Instead, enzymes feature a larger variety of sites, which evolved to guarantee sufficient stabilization of both the oxidation states of copper, which have quite different chemical properties. All these sites use at least one histidine to bind copper together with other ligands such as cysteine and methionine or, less commonly, glutamate and aspartate.[12] Copper proteins are less pervasive than zinc and nonheme iron proteins and typically account for less than 1% of an organism's proteome. Whereas archaeal and bacterial copper-proteins most often have homologues in all the domains of life, eukaryotes contain a significant share of eukaryotic-specific copper proteins. These proteins are mainly copper-dependent oxidoreductases, whose number could be higher than that in less complex organisms because of the compartmentalization of the eukaryotic cell. In agreement with this, a parallel expansion of copper ion transporters in eukaryotes is also observed, albeit less important in numerical terms.

There are several pathways or multiprotein machineries that interact with metals or metal cofactors transiently, for example, for the synthesis and assembly of complex cofactors or for their transport from outside or within the cell. Heme is a notable example in this context, because it requires relatively complex machineries for its biosynthesis, its insertion into hemoproteins of the *c*-type, and its uptake from external sources. On the other hand, heme-binding patterns are quite simple. It is possible to apply the methods previously described also to the analysis of complicated systems such as the above ones by supporting the search for the relevant protein domains and, when possible, for the corresponding binding patterns with the investigation of the co-occurrence of the various genes required for the process under study. In this way, it is possible to obtain information on the evolution of the processes,[24] on which organisms are able to perform which process, or on the possible existence of uncharacterized enzymes along the pathways of interest.[25] The different solutions developed by Nature to the same problem can be identified (e.g., the variability of heme-binding modes within a common protein scaffold[25]). Monoheme cytochromes *c* occur in all three domains of life. A soluble cytochrome *c* and a cytochrome *c* domain within the membrane-bound cytochrome bc_1 complex are present in the large majority of eukaryotes. The number and functional diversity of monoheme cytochromes *c* is somewhat larger in bacteria, where they are involved in various respiratory processes as well as in metabolic or biosynthetic processes as electron carriers. On the other hand, monoheme cytochromes *c* are rare in archaea, where again they presumably serve as electron carriers in a limited set of respiratory processes. Regarding heme biosynthesis and uptake, it was observed that different systems exist for prokaryotic organisms belonging to different branches of the tree of life.[25] Some prokaryotes presumably cannot perform either of the two processes (14%), some can perform only one of them (40%), and some can perform both of them (46%). A large share of Gram-positive pathogens do perform heme uptake from the host, suggesting that this process can be a potential target for wide-spectrum antibiotics.[25]

Caveats and Solutions

The bioinformatic approaches developed for the prediction of metalloproteomes ultimately rely on the data stored in the PDB, which is the unique source for information on the metal-binding sites in proteins. As a consequence, our predictions should improve with the increasing number of metalloproteins deposited in the PDB. As of June 2008, 9797 of the 51261 structures deposited in the PDB bind at least one essential transition metal ion. Based on literature mining and on the classification of the protein structures in the CATH[26] and SCOP[27] databases, about 85% (8313) of these interactions are physiologically relevant. Hence, 16% of the PDB structures are of a protein binding a transition metal ion (Figure S2-A, Supporting Information). Figure 5A, which was drawn for this review as described in the Supporting Information, shows that the number of structures being deposited in the PDB every year has been increasing constantly in the last seven years. This trend reflects both the progress in the experimental techniques to solve macromolecular structures and the efforts of Structural Genomics projects. The fraction of metalloproteins binding a transition metal ion has been constantly about 16% of all new structures, so their number has also been steadily increasing. However, the number of new protein families,[27] which typically correspond to an individual Pfam profile, deposited in the PDB every year has been rapidly decreasing since 2004 (Figure 5B), implying that the new structures are mainly of proteins belonging to families already structurally characterized. This number does not suffer from the high redundancy of the PDB and is thus more informative. Even if less rapidly, also the number of new metalloprotein families has been decreasing since 2004. Currently the PDB contains 3464 distinct families, according to SCOP,[27] 580 of which include at least one metalloprotein structure (17% of the total), as shown in Figure S2-B, Supporting Information. The decreasing rate of discovery of new protein families implies that the above numbers should remain fairly stable in time, and thus the results of metalloproteome searches should also be correspondingly stable. Note that the above considerations apply to soluble, globular proteins. Membrane proteins or intrinsically unfolded proteins cannot be handled systematically at present due to the paucity of experimental data available for them.

A limitation of the protocols described here is their low selectivity for the detection of metal-binding sites at the interface of two or more chains. This is due to the simplicity of the MBPs on each chain, often composed of only two ligands per chain, making it difficult to remove false positives. Similar con-

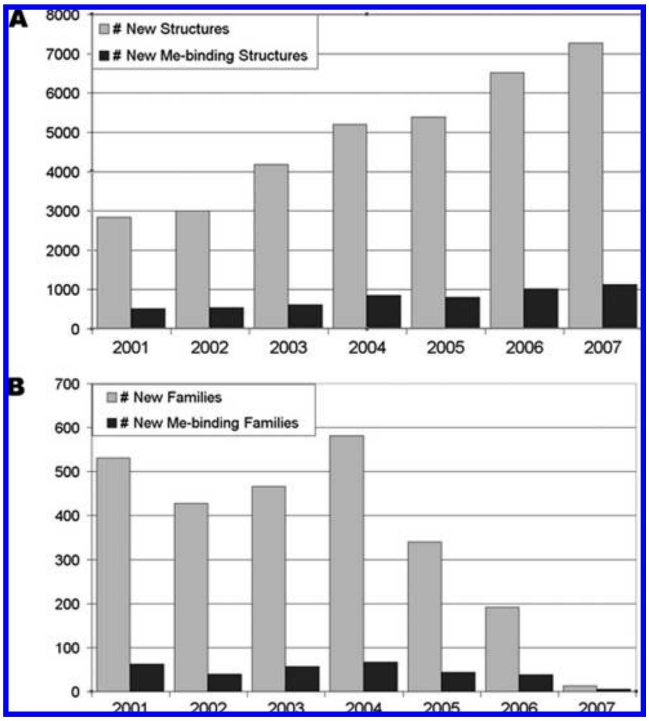

FIGURE 5. Distribution of (A) structures and (B) SCOP[27] families being deposited in the PDB per year over the last 7 years (gray) as of June 2008. The corresponding numbers for metalloproteins are reported in black. The gray bars of Figure 5A were built using the data reported on the PDB Web site (http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100), whereas the black bars were obtained on the basis of the release date of the PDB structures of metalloproteins (the ensemble of metalloproteins was created as reported in the methods section). The gray bars of Figure 5B were obtained on the basis of the release date of new families, as defined in the SCOP database. In this classification system, families include proteins that are clearly evolutionarily related; consequently, the pairwise sequence identity between any pair of proteins within the same family is usually at least 30%.[26] The parseable file in http://scop.mrc-lmb.cam.ac.uk/scop/parse/dir.cla.scop.txt_1.73 was used to associate PDB structures with families. The black bars of Figure 5B were built by applying the same process to the ensemble of metalloproteins.

siderations explain the difficulty of predicting metalloproteins binding nontransition metal ions or metal-binding cofactors, such as heme. Metal ions that belong to the s- and p-series are much harder (in the sense of the hard—soft theory) than transition metal ions. They thus have preferences for oxygen donors, which can be provided by the protein backbone rather than side chains, so their MBPs are most often less specific. Analogously, some proteins bind complex metal-binding cofactors through nonspecific hydrophobic interactions. There are some possible ways to tackle this problem. Domain recognition will work independently of the identity of the metal ion/metal-binding cofactor required by the metalloprotein, and thus there is no particular difference in using domain searches for nontransition rather than transition metal ions. The problem of domain searches alone is, as mentioned, their relatively low precision. In the case of transition metal ions, this was circumvented by the use of MBPs as a filter. For those nontransition metals that are typically bound to at least a pair of amino acid side chains, such as calcium-binding proteins, the filter could be modified by using structure prediction methods to obtain an indication of whether suitable side chains are close enough to one another in the structure. Alternatively, one could rely on the introduction of new criteria based on the properties of the second coordination sphere in addition to or in place of MBP filtering. This would require that the information on the second coordination sphere, which must be obtained from structures, is translated into sequence motifs.

Concluding Remarks

Metalloproteomics is an increasingly important area of research within systems biology approaches.[5,28] To achieve biological insights, researchers require a synergistic combination of data. To this end, the identification of metal cofactors in a protein can be extremely useful for its functional assignment, as well as to place it in the proper cellular context. In this Account, we showed how bioinformatics methods can be applied in searching for metalloproteomes by exploiting the available knowledge in the PDB and in domain libraries.

There is still significant room for expanding the portfolio of computational resources of use in metalloproteomics and, more generally, in biological inorganic chemistry. In this context, a future advancement is the development of databases dedicated to store functional and structural information on metal ions in metalloproteins. Despite the copious number of available databases aimed at providing the scientific community with tools for the investigation of biomolecules, very few of these resources have been dedicated to the analysis of metal ions in proteins, and the most important of them have been discontinued.[29,30] The lack of this kind of resources contributes to making biological inorganic chemistry a specialized area of study, poorly accessible to nonexpert scientists, and discourages its study. At present, the quality of the results that can be obtained through computational methods largely depends on the quality of the manual work spent to select the starting metallo-data sets (metal-binding domains, metal-binding structures, and so on). Such work typically relies on literature analysis, which in some cases may leave a degree of uncertainty on the actual metal-binding properties of proteins in vivo. For example, CHCH domains have been reported to bind copper based on biochemical data, yet the first structurally characterized protein containing this domain did not bind copper, raising doubts on whether CHCH domains generally bind copper in vivo. Therefore, there is also the need for creating methods and automated tools for the analysis and classification of metal-binding sites on the basis of objective criteria. The combination of these kinds of tools with the above-mentioned databases could lead to the development of resources addressing a larger number of proteomes and automatically maintained up to date.

Another area of future work points to the analysis of regulation by metal sensors.[31] Metal sensors, indeed, regulate metalloprotein expression by binding to DNA promoter regions; these regions are similar in co-regulated sets of genes. The finding of the complete set of these DNA recognition sequences could thus provide the complete set of metallo-regulated operons (metalloregulons). The analysis of metalloregulons would lead to the identification of new metalloproteins, potentially with new metal binding sites or domains. Significant efforts are also needed to integrate the metalloproteome analyses with other data sets. As an example, the integration of metalloproteomics with transcriptomics and interactomics data could be especially useful in the comprehension of metal ion homeostasis, an area of research of particular importance given that homeostasis impairment is often associated with important diseases.[32−34]

Supporting Information Available. Figure S1 showing the number of putative zinc-binding proteins as a function of the proteome size, Figure S2 showing percentages of transition metal-binding (A) proteins and (B) SCOP[27] families in the PDB, supplementary methods describing how Figure 5 was built, Table S1 listing Pfam metal-binding domains (zinc, copper, nonheme iron), Table S2 listing organisms analyzed, Table S3 listing putative bacterial metalloproteins grouped by organism, Table S4 listing putative archeal metalloproteins grouped by organism, and Table S5 listing putative eukaryotic metalloproteins grouped by organism (the fields of Tables S3−S5 contain (i) NCBI code, (ii) protein length, (iii) brief description as reported in the proteome release, (iv) potential zinc-binding pattern(s) in the sequence, and (v) domain composition (when a domain is followed by a pattern within brackets, the pattern is localized within the domain)). This material is available free of charge via the Internet at http://pubs.acs.org.

BIOGRAPHICAL INFORMATION

Claudia Andreini was born in Florence in 1977. She obtained the Italian degree of Doctor in Chemistry at the University of Florence (Italy) in 2002 and then obtained her Ph.D. degree in Chemical Sciences at the same university in 2006. In her Ph.D. studies, she focused on the characterization of metalloproteins through bioinformatic approaches, acquiring skills both on the use and the development of databases and bioinformatic tools. Now she has a postdoctoral position at the Magnetic Resonance Center (University of Florence). She is collaborating with the Istituto Superiore di Sanità in Rome to model protein−protein interactions in a project for the development of new-generation anti-AIDS vaccines and with the European Bioinformatics Institute (EBI) in Cambridge to annotate and classify metal sites in protein structures.

Ivano Bertini was born in Florence in 1940; he obtained the Italian degree of Doctor in Chemistry at the University of Florence (Italy) in 1964 and then the Libera Docenza in 1969. He became full professor in General and Inorganic Chemistry in 1975 at the University of Florence, where he is now. He received the Laurea Honoris Causa from the University of Stockholm in 1998, Ioannina in 2002, and Siena in 2003. He is a physical bioinorganic chemist who has cultivated NMR since 1965. Since 1974, he has been working with metals in life sciences. He has solved the solution structure of more than 100 proteins, has discussed the mobility of some of them, and has studied a number of protein−protein interactions. He is founder and director of the Magnetic Resonance Center (CERM), an NMR-based infrastructure operating in the field of life sciences.

Antonio Rosato was born in Florence in 1971. He obtained the Italian degree of Doctor in Chemistry in 1995 and then his Ph.D. degree in 1998, both at the University of Florence (Italy). He is now Associate Professor of Chemistry at the same University. His research interests focus on the theory and practice of NMR spectroscopy applied to paramagnetic systems and on the application and development of bioinformatic tools to browse gene banks and genome sequences, with a particular interest in metalloproteins.

FOOTNOTES

[*] To whom correspondence should be addressed. E-mail: ivanobertini@cerm.unifi.it. Phone: + 39 055 4574270. Fax: + 39 055 4574271.

[†] E-mail addresses: andreini@cerm.unifi.it; rosato@cerm.unifi.it.

REFERENCES

1 Nielsen, F. H. Evolutionary Events Culminating in Specific Minerals Becoming Essential for Life. *Eur. J. Nutr.* **2000**, *39*, 62–66.

2 Bertini, I.; Sigel, A.; Sigel, H. *Handbook on Metalloproteins*; Marcel Dekker: New York, 2001; pp 1−1800.

3 Andreini, C.; Bertini, I.; Cavallaro, G.; Holliday, G. L.; Thornton, J. M. Metal Ions in Biological Catalysis: From Enzyme Databases to General Principles. *J. Biol. Inorg. Chem.* **2008**, *13*, 1205–1218.

4 Ideker, T.; Galitski, T.; Hood, L. A New Approach to Decoding Life: Systems Biology. *Annu. Rev. Genomics Hum. Genet.* **2001**, *2:343−72.*, 343–372.

5 Shi, W.; Chance, M. R. Metallomics and Metalloproteomics. *Cell. Mol. Life Sci.* **2008**, *65*, 3040–3048.

6 Finn, R. D.; Mistry, J.; Schuster-Bockler, B.; Griffiths-Jones, S.; Hollich, V.; Lassmann, T.; Moxon, S.; Marshall, M.; Khanna, A.; Durbin, R.; Eddy, S. R.; Sonnhammer, E. L.; Bateman, A. Pfam: Clans, Web Tools and Services. *Nucleic Acids Res.* **2006**, *34*, D247–D251.

7 Andreini, C.; Banci, L.; Bertini, I.; Rosato, A. Zinc through the Three Domains of Life. *J. Proteome Res.* **2006**, *5*, 3173–3178.

8 Eddy, S. R. Profile Hidden Markov Models. *Bioinformatics* **1998**, *14*, 755–763.

9 Andreini, C.; Bertini, I.; Rosato, A. A Hint to Search for Metalloproteins in Gene Banks. *Bioinformatics* **2004**, *20*, 1373–1380.

10 Andreini, C.; Banci, L.; Bertini, I.; Rosato, A. Counting the Zinc Proteins Encoded in the Human Genome. *J. Proteome Res.* **2006**, *5*, 196–201.

11 Andreini, C.; Banci, L.; Bertini, I.; Elmi, S.; Rosato, A. Non-Heme Iron through the Three Domains of Life. *Proteins* **2007**, *67*, 317–324.

12 Andreini, C.; Banci, L.; Bertini, I.; Rosato, A. Occurence of Copper through the Three Domains of Life: A Bioinformatic Approach. *J. Proteome Res.* **2008**, *1*, 209–216.

13 Lin, H. H.; Han, L. Y.; Zhang, H. L.; Zheng, C. J.; Xie, B.; Cao, Z. W.; Chen, Y. Z. Prediction of the Functional Class of Metal-Binding Proteins from Sequence Derived Physicochemical Properties by Support Vector Machine Approach. *BMC Bioinf.* **2006**, *7*, Suppl 5: S13.

14 Passerini, A.; Andreini, C.; Menchetti, S.; Rosato, A.; Frasconi, P. Predicting Zinc Binding at the Proteome Level. *BMC Bioinf.* **2007**, *5*, 8–39.

15 Ebert, J. C.; Altman, R. B. Robust Recognition of Zinc Binding Sites in Proteins. *Protein Sci.* **2008**, *17*, 54–65.

16 Shu, N.; Zhou, T.; Hovmoller, S. Prediction of Zinc-Binding Sites in Proteins from Sequence. *Bioinformatics* **2008**, *24*, 775–782.

17 Vallee, B. L.; Auld, D. S. Active-Site Zinc Ligands and Activated H_2O of Zinc Enzymes. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 220–224.

18 Rawlings, N. D.; Morton, F. R.; Kok, C. Y.; Kong, J.; Barrett, A. J. MEROPS: The Peptidase Database. *Nucleic Acids Res.* **2008**, *36*, D320–D325.

19 Malgieri, G.; Russo, L.; Esposito, S.; Baglivo, I.; Zaccaro, L.; Pedone, E. M.; Di Blasio, B.; Isernia, C.; Pedone, P. V.; Fattorusso, R. The Prokaryotic Cys2His2 Zinc-Finger Adopts a Novel Fold as Revealed by the NMR Structure of *Agrobacterium tumefaciens* Ros DNA-Binding Domain. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 17341–17346.

20 Krishna, S. S.; Majumdar, A.; Grishin, N. V. Structural Classification of Zinc Fingers: Survey and Summary. *Nucleic Acid Res.* **2003**, *31*, 532–550.

21 O'Halloran, T. V.; Culotta, V. C. Metallochaperones: An Intracellular Shuttle Service for Metal Ions. *J. Biol. Chem.* **2000**, *275*, 25057–25060.

22 Rae, T. D.; Schmidt, P. J.; Pufahl, R. A.; Culotta, V. C.; O'Halloran, T. V. Undetectable Intracellular Free Copper: The Requirement of a Copper Chaperone for Superoxide Dismutase. *Science* **1999**, *284*, 805–808.

23 Banci, L.; Rosato, A. Structural Genomics of Proteins Involved in Copper Homeostasis. *Acc. Chem. Res.* **2003**, *36*, 215–221.

24 Bertini, I.; Cavallaro, G.; Rosato, A. Evolution of Mitochondrial-Type Cytochrome *c* and of the Protein Machinery for Their Assembly. *J. Inorg. Biochem.* **2007**, *101*, 1798–1811.

25 Cavallaro, G.; Decaria, L.; Rosato, A. Genome-Based Analysis of Heme Biosynthesis and Uptake in Prokaryotic Systems. *J. Proteome Res.* **2008**, *11*, 4946—4954.

26 Greene, L. H.; Lewis, T. E.; Addou, S.; Cuff, A.; Dallman, T.; Dibley, M.; Redfern, O.; Pearl, F.; Nambudiry, R.; Reid, A.; Sillitoe, I.; Yeats, C.; Thornton, J. M.; Orengo, C. A. The CATH Domain Structure Database: New Protocols and Classification Levels Give a More Comprehensive Resource for Exploring Evolution. *Nucleic Acids Res.* **2007**, *35*, D291–D297.

27 Andreeva, A.; Howorth, D.; Chandonia, J. M.; Brenner, S. E.; Hubbard, T. J.; Chothia, C.; Murzin, A. G. Data Growth and Its Impact on the SCOP Database: New Developments. *Nucleic Acids Res.* **2008**, *36*, D419–D425.

28 Bertini, I.; Rosato, A. Bioinorganic Chemistry in the Post-Genomic Era. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3601–3604.

29 Degtyarenko, K. N.; North, A. C. T.; Findlay, J. B. C. PROMISE: A Database of Bioinorganic Motifs. *Nucleic Acids Res.* **1999**, *27*, 233–236.

30 Castagnetto, J. M.; Hennessy, S. W.; Roberts, V. A.; Getzoff, E. D.; Tainer, J. A.; Piquet, M. E. MDB: The Metalloprotein Database and Browser at The Scripps Research Institute. *Nucleic Acids Res.* **2002**, *30*, 379–382.

31 Tottey, S.; Harvie, D. R.; Robinson, N. J. Understanding How Cells Allocate Metals Using Metal Sensors and Metallochaperones. *Acc. Chem. Res.* **2005**, *38*, 775—783.

32 Lutsenko, S.; Barnes, N. L.; Bartee, M. Y.; Dmitriev, O. Y. Function and Regulation of Human Copper-Transporting ATPases. *Physiol. Rev.* **2007**, *87*, 1011–1046.

33 Rouault, T. A.; Tong, W. H. Iron-Sulfur Cluster Biogenesis and Human Disease. *Trends Genet.* **2008**, *24*, 398–407.

34 Shoubridge, E. A. Cytochrome c Oxidase Deficiency. *Am. J. Med. Genet.* **2001**, *106*, 46–52.