Metal binding in proteins: machine learning complements X-ray absorption spectroscopy

Marco Lippi[1], Andrea Passerini[2], Marco Punta[3], and Paolo Frasconi[4]

[1] Dipartimento di Ingegneria dell'Informazione, Università di Siena, Italy
[2] Dipartimento di Ingegneria e Scienza dell'Informazione, Università di Trento, Italy
[3] Wellcome Trust Sanger Institute, Hinxton, UK
[4] Dipartimento di Sistemi e Informatica, Università di Firenze, Italy

Abstract. We present an application of machine learning algorithms for the identification of metalloproteins and metal binding sites on a genome scale. An extensive evaluation conducted in combination with X-ray absorption spectroscopy shows the great potentiality of the approach.

1 Metal binding in proteins

A significant fraction of known proteins is believed to bind metal ions in their native conformation. Metal ions play a variety of crucial roles in proteins [1], from stabilizing their three dimensional structure, to acting as cofactors in enzyme catalysis. Moreover, metals are implicated in many diseases for which medicine is still seeking an effective treatment, such as Parkinson's or Alzheimer's [2] . Identifying unknown metalloproteins and detecting their metal binding site(s) is an important step in understanding their function and characterizing many crucial processes involved in living systems. A metal binding site is characterized through its metal ion, the protein amino acid residues directly involved in binding it (called ligands) and the binding geometry, i.e. the spatial arrangement of the ion and its ligands. Some metal binding sites actually involve compounds (e.g. the heme group binding hemoglobin) including one or more ions, and some proteins contain more than one metal binding site. The problem of identifying and characterizing metalloproteins can be seen as a series of increasingly complex tasks. Given a protein, one aims at: 1) determining whether it is a metalloprotein or not, i.e. if it binds metal ions in its native conformation; 2) determining the metal bonding state of each of its residues, i.e. whether they bind a metal ion or not; 3) determining the composition of metal binding sites, i.e. the number of ions binding the protein and the set of their respective ligands. Answers may be obtained experimentally, by means of in-silico prediction tools, or by a combination of the two classes of methods.

Experimental methods. High-throughput techniques based on X-ray absorption spectroscopy [3] (HT-XAS) allow detection, identification and quantification of metals bound to proteins based on the energy and intensity of the X-ray fluorescence signal emitted by the metals. HT-XAS can thus be used to address the first task, i.e. metalloprotein identification. Exact determination of binding sites, however, is only possible using more labor-intensive techniques, such as X-ray

crystallography or Nuclear Magnetic Resonance, which provide high-resolution three-dimensional structural information. Even when a 3D structure is available, exact characterization of binding sites may be non trivial and error prone. For example false positives may be due to spurious artifacts where metals bind at adventitious sites, and false negatives may emerge when metalloproteins are experimentally solved in their apo-form lacking the metal ion.

In silico methods. The three above tasks can be tackled from a machine learning point of view. Here we focus on predictions from sequence alone[1], where (1) is a sequence classification problem, (2) is a sequence labeling problem, and (3) is a more complex structured output problem from sequences to bipartite graphs. Some simplifying assumptions may be made to reduce the difficulty of these problems in their generality. First, prediction may be limited to transition metals and a small number of candidate residues (CYS and HIS). Transition metals (especially iron and zinc) are the most commonly found ions in proteins, covering about 2/3 of all known metalloproteins. Their preferred ligands are CYS and HIS, followed by ASP and GLU, which have a much lower binding propensity given their relatively high abundance in proteins. Finally, the solution space may be limited to sites where an ion is coordinated by four or less residues as more complex binding sites are extremely rare.

MetalDetector. The MetalDetector software [5, 6] uses state-of-the-art machine learning methods to solve the above three prediction problems. A first version of the software [5] employed Disulfind predictor to identify cysteine disulfide bridges [7] and a combination of support vector machines and bidirectional recurrent neural networks for metal bonding state prediction. The current version[2] of the server [6] employs a two-stage approach for metal bonding state and metal binding sites prediction respectively. The first stage relies on an SVM-HMM [8] which collectively assigns the bonding state of all the CYS/HIS residues in the sequence. Residues predicted as metal-bound are fed to the second stage. Here a search-based structure output approach greedily adds links between candidate ligands and candidate ions, until each ligand is connected to an ion. The search is guided by a kernel-based scoring function trained to score correct moves higher than incorrect moves. The problem has the structure of a weighted matroid, which is basically the discrete counterpart of concave functions. The greedy search is thus guaranteed to lead to the global optimum of the (learned) scoring function. The method was initially introduced in [9] and further refined and analyzed in [10], where an extensive experimental validation across different protein structural folds and superfamilies was conducted.

Being able to address all three predictions problems, MetalDetector is a natural candidate to complement information provided by high-throughput experimental techniques like HT-XAS. The potential impact of this integration was recently shown on a large-scale experiment aimed at identifying potential metalloproteins within the New York SGX Research Center for Structural Genomics.

[1] For applications of machine learning techniques to 3D structure data see e.g. [4]

[2] MetalDetector is available as a web server at `http://metaldetector.dsi.unifi.it`.

2 Results

MetalDetector and HT-XAS were jointly employed in a recent study [11] in order to identify metal-bonded residues in 3,879 purified proteins generated by the New York SGX Research Center for Structural Genomics and belonging to hundreds of different protein families.

Of the whole set of proteins, 343 were identified by HT-XAS to contain at least one metal ion among Mn, Fe, Co, Ni, Cu and Zn. The experimental analysis described in [11] compares the level of agreement between MetalDetector and HT-XAS predictions: to this aim, MetalDetector predictions at residue level have been combined in order to define a protein score, which is used to predict whether that protein is a metalloprotein or not. This level of agreement obviously depends on the aggregation criterion used to produce such protein scores for MetalDetector: in these experiments, the adopted criterion was to predict a protein to be a metalloprotein if for at least N residues (either CYS or HIS) the probability of metal bonding state, as predicted by MetalDetector, exceeded a certain threshold T_M. By choosing different values for N and T_M, different predictions can be accordingly obtained: the experiments showed that, at the same recall level (i.e., when MetalDetector predicts the same number of metalloproteins as HT-XAS), MetalDetector and HT-XAS agree from 32% up to 45% of the cases, depending on the choice of N and T_M (note that a random baseline predictor would achieve a 10% of precision with respect to the metalloproteins identified by HT-XAS).

In addition, it must be underlined that in many protein samples metal occupancy can be low, and therefore metal atoms cannot be detected by HT-XAS. MetalDetector can in these cases complement HT-XAS evidence by suggesting potentially missed metalloproteins. This happens, for example, for proteins 11211f and 11213j, which share the Pfam SCO1/SenC domain (PF02630) involved in biogenesis of respiratory and photosynthetic systems. Protein 11211f shares 26% sequence identity with Human SCO2 protein, a mitochondrial membrane-bound protein involved in copper supply for the assembly of cytochrome c oxidase. The residues predicted by MetalDetector to bind metal in 11211f align to the residues that in the NMR structure of Human SCO2 bind a Cu^+ ion. This is likely one of the cases where the HT-XAS method fails in identifying a metalloprotein, while MetalDetector not only seems to recover the false negative, but also to correctly predict the position of the binding site.

A comprehensive approach combining the use of MetalDetector predictions with homology modeling has also been object of analysis and showed that the proposed computational methodology can represent an extremely powerful tool for the study of metalloproteins.

3 Perspectives

Recent years have witnessed a dramatic increase in the availability of high-throughput experimental techniques for the analysis of biological data. This scenario provides an unprecedented opportunity for machine learning approaches

to deal with large amount of data and continuously novel problems and challenges. Structural genomics, which aims to map the protein sequence space with structural information, is indeed pursuing a tight integration of high-throughput experimental techniques and modeling approaches. Characterization of metalloproteins is an interesting example of how experimental techniques and machine learning approaches can be fruitfully combined to deepen our understanding of biological systems.

4 Acknowledgments

ML, AP and PF were partially supported by grant PRIN 2009LNP494 (Statistical Relational Learning: Algorithms and Applications) from Italian Ministry of University and Research. MP is supported by Wellcome Trust (grant numbers WT077044/Z/05/Z). We gratefully acknowledge the collaboration with W. Shi, J. Bohon, M. Sauder, R. D'Mello, M. Sullivan, J. Toomey, D. Abel, S.K. Burley, B. Rost and M.R. Chance.

References

1. Bertini, I., Sigel, A., Sigel, H.: Handbook on Metalloproteins. M. Dekker (2001)
2. Barnham, K.J., Bush, A.I.: Metals in Alzheimer's and Parkinson's diseases. Current Opinion in Chemical Biology **12**(2) (2008) 222–228
3. Shi, W., Zhan, C., Ignatov, A., Manjasetty, B.A., Marinkovic, N., Sullivan, M., Huang, R., Chance, M.R.: Metalloproteomics: High-throughput structural and functional annotation of proteins in structural genomics. Structure **13**(10) (2005) 1473–1486
4. Babor, M., Gerzon, S., Raveh, B., Sobolev, V., Edelman, M.: Prediction of transition metal-binding sites from apo protein structures. Proteins **70**(1) (2007) 208–217
5. Lippi, M., Passerini, A., Punta, M., Rost, B., Frasconi, P.: Metaldetector: a web server for predicting metal-binding sites and disulfide bridges in proteins from sequence. Bioinformatics **24**(18) (2008) 2094–2095
6. Passerini, A., Lippi, M., Frasconi, P.: Metaldetector v2.0: predicting the geometry of metal binding sites from protein sequence. Nucl. Ac. Res. **39**(Web-Server-Issue) (2011) 288–292
7. Ceroni, A., Passerini, A., Vullo, A., Frasconi, P.: Disulfind: a disulfide bonding state and cysteine connectivity prediction server. Nucl. Ac. Res **34** (2006) 177–181
8. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. J. Mach. Learn. Res. **6** (2005) 1453–1484
9. Frasconi, P., Passerini, A.: Predicting the geometry of metal binding sites from protein sequence. In: NIPS 21, Vancouver, Canada, MIT Press (2008) 465–472
10. Passerini, A., Lippi, M., Frasconi, P.: Predicting metal-binding sites from protein sequence. IEEE/ACM Trans. Comput. Biology Bioinform. **9**(1) (2012) 203–213
11. Shi, W., Punta, M., Bohon, J., Sauder, J.M., D'Mello, R., Sullivan, M., Toomey, J., Abel, D., Lippi, M., Passerini, A., Frasconi, P., Burley, S.K., Rost, B., Chance, M.R.: Characterization of metalloproteins by high-throughput x-ray absorption spectroscopy. Genome Research **21**(6) (2011) 898–907