Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw238
Applications Note

Sequence analysis

MetalPredator: a web server to predict iron–sulfur cluster binding proteomes

Yana Valasatava[1], Antonio Rosato[1,2], Lucia Banci[1,2] and Claudia Andreini[1,2,*]

[1]Magnetic Resonance Center (CERM) and [2]Department of Chemistry, University of Florence, 50019 Sesto Fiorentino, Italy

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Abstract

Motivation: The prediction of the iron–sulfur proteome is highly desirable for biomedical and biological research but a freely available tool to predict iron–sulfur proteins has not been developed yet.
Results: We developed a web server to predict iron–sulfur proteins from protein sequence(s). This tool, called MetalPredator, is able to process complete proteomes rapidly with high recall and precision.
Availability and Implementation: The web server is freely available at: http://metalweb.cerm.unifi.it/tools/metalpredator/.
Contact: andreini@cerm.unifi.it
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Metalloproteins are widespread in living organisms and essential for many biological processes (Andreini *et al.*, 2008). Misregulation of levels and usage of metal ions in cells is associated with major diseases. Thus, identifying potential metalloproteins is a crucial task not only for the functional characterization of proteins but also for applications in biomedical research and drug design. Furthermore, knowledge on the complete ensemble of proteins that bind a specific metal, i.e. each specific metalloproteome, permits the study of how organisms evolved to adapt to different metal availability in their environments. The experimental identification of metal-binding sites can be quite difficult and costly, especially when attempted at the whole proteome scale. Consequently, various bioinformatics approaches have been developed to predict the metal-binding sites in a single sequence (Andreini *et al.*, 2004; Gladyshev and Zhang, 2013; Lin *et al.*, 2006; Passerini *et al.*, 2011) but they do not allow metalloproteomics data analyses. Some of the authors of the present article developed an initial prototype of a tool to predict metalloproteomes (Andreini *et al.*, 2011), which however lacked a user-friendly interface.

We propose a novel computational tool, called *MetalPredator*, to predict metal-binding sites in protein sequence(s) at the whole proteome scale. The tool integrates an existing domain-based approach (Andreini *et al.*, 2011) with a new one designed to search for metal-binding motifs found in proteins with known structure. MetalPredator uniquely combines global and local searches to define whether a protein is a potential metalloprotein.

To validate our general methodology, here we focused on the prediction of the iron–sulfur (Fe–S) proteome, which has recently been the subject of an extensive body of experimental work worldwide (Andreini *et al.*, 2016; Paul and Lill, 2015). Fe–S clusters are among the most ancient biological metal cofactors and proteins that bind them (iron–sulfur proteins, ISP hereafter) play crucial roles in many cellular processes. MetalPredator is, to our knowledge, the only available tool that performs Fe–S protein/proteome prediction via a web interface. It featured a very good performance in terms of precision and recall. MetalPredator is available at http://metalweb.cerm.unifi.it/tools/metalpredator/.

2 Methods

To identify metal-binding sites in protein sequences, MetalPredator uses two libraries of Hidden Markov Model profiles that represent (1) Pfam domains and (2) structural motifs binding Fe–S clusters. Metal-binding motifs are defined by splitting the Minimal

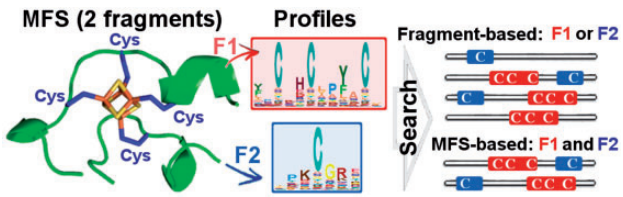

Fig. 1. Pipeline for the creation of library (2) and to search ISPs in agreement with criteria (C) and (D)

Functional Sites (MFSs) stored in MetalPDB (Andreini *et al.*, 2013) into fragments. Each fragment is a continuous stretch of protein sequence containing at least one metal ligand. The library (1) was built as described in (Andreini *et al.*, 2011). It contains the profiles of both Pfam domains for which the Fe–S cluster ligands are known and domains annotated as Fe–S cluster binding but lacking information on the ligands. To build the library (2) each ISP in MetalPDB was compared to UniRef50 representative sequences using PSI-Blast. All the hits with conserved ligands were used to build a sequence profile. From this profile we extracted the profiles of the distinct fragments corresponding to the MFS(s) in the initial input sequence.

MetalPredator uses the hmmscan tool (Eddy, 1998) to match every input sequence to the profiles contained in the libraries. An input sequence is identified as a potential ISP if at least one of these conditions applies:

(A) The profile of a domain with associated ligands (library 1) matches the sequence with an e-value lower than 10^{-5} and ligands are conserved in the sequence (*Domain search with pattern filter*).
(B) The profile of a domain with no information on ligands available (library 1) matches the sequence with an e-value lower than 10^{-7} (*Domain search without pattern filter*).
(C) All fragment profiles of a given MFS (library 2) match the sequence with an e-value lower than 10^{-3} and the corresponding ligands are conserved in the sequence (*MFS search*, Fig. 1).
(D) At least one fragment profile of a given MFS (library 2) matches the sequence with an e-value lower than 10^{-3} and the corresponding fragment ligands are conserved in the sequence (*Fragment search*, Fig. 1).

We adjusted the above e-value thresholds to have the best $F_{0.5}$ measure for each criterion (see Supplementary Table S1). This measure weighs precision more than recall, thus reducing the number of false positives in the output as well as the experimental effort needed to validate the predictions. For the calibration we used two datasets of sequences (positives and negatives), taken from a subset of the Protein Data Bank (Berman *et al.*, 2000) filtered at a sequence identity level of 25% (PDB25). The positive dataset included all 163 ISPs in PDB25. The negative dataset included 2607 PDB25 sequences that are not homologous to ISPs, and either bind a metal ion different from iron in MetalPDB or are not metalloproteins but contain at least four Cys.

3 Results

3.1 Performance of the tool

To avoid overfitting, we assessed MetalPredator by using a leave-one-out cross-validation (LOOCV) approach on the entire PDB25. In LOOCV each training set is created by taking all the samples except one, and the test set is the sample left out. The procedure is repeated by creating as many training and test sets as are the samples available. Accordingly, the precision and recall of MetalPredator are 89.1 and 75.5%, respectively (Table 1).

Table 1. Performance of MetalPredator

Test set	Training set	TP	FP	FN	Precision (%)	Recall (%)
LOOCV	PDB25-1	123	15	40	89.1	75.5
E. coli	PDB25*	123	20	26	86.0	82.5
E. coli	PDB70*	132	23	17	85.2	88.6

*Structures of *E. coli* proteins were excluded from the training set. All datasets are available from the MetalPredator site.

As a further assessment, we repeated the training of MetalPredator excluding all sequences of *Escherichia coli* proteins and then predicted the Fe–S proteome of this organism (Table 1 and Supplementary Table S2). A very similar approach was used by (Estellon *et al.*, 2014), using PDB70 as the starting dataset. Our tool has similar precision (85.2% versus 86.5) and higher recall (88.6% versus 66.2%) than the 'Mixed model' of (Estellon *et al.*, 2014). Notably, it was possible to build a 3D model for seven of our false positives, based on homology modeling, showing that they contain a plausible binding site for a Fe–S cluster (see Supplementary Table S3). This would make the precision of MetalPredator even higher.

3.2 Description of the web interface

The *home page* allows users to submit the query protein sequence(s) in FASTA format by either pasting them directly or uploading a file. It is possible to provide an e-mail address to receive a notification when the results are ready. The output is displayed in a *Summary page*, which gives an overview of all predicted ISPs. The report contains the results of all predictions based on the criteria A–D described in the Section 2 in separate columns. Potential ISPs are ordered according to the number of methods that supported the prediction. The column corresponding to each criterion reports the location of the putative Fe–S cluster binding site. It is possible to click on the site to visualize the details of each prediction. Only for criterion B, the column reports the Fe–S cluster-binding domain identified. A downloadable csv file is created upon user's selection of the columns to be included. Additionally, the interface presents the results for each method on separate tabs. In each tab, the predicted ISPs are sorted by e-value. Note that there can be multiple predictions for the same sequence because more than one profile can be matched with a score better than the selected thresholds. The *Details* link pops up the list of all profiles matching a given sequence with additional information on the match. *Filter* and *Download* options are available to facilitate the analysis of the output.

4 Concluding remarks

The prediction of metalloproteomes is highly desirable for systems biology and biomedical research efforts. Our tool integrates an existing methodology for domain-based predictions (Andreini *et al.*, 2011) and an approach to search for metal-binding motifs, based on MFSs or fragments thereof. This integration exploits the complementarity between the global properties of protein domains and the local nature of MFSs. In the present paper we used our approach to predict iron–sulfur proteins but, in principle, the methodology can be straightforwardly applied to search for other groups of metalloproteins. MetalPredator can process the entire proteome of any organism in minutes to a few hours (e.g. for the human proteome), and thus can be applied to any newly sequenced organism, including eukaryotes.

Funding

This work was supported by CIRMMP and by the European Commission through the BioMedBridges and EGI-Engage project (grants nos. 284209 and 654142).

Conflict of Interest: none declared.

References

Andreini,C. *et al*. (2004) A hint to search for metalloproteins in gene banks. *Bioinformatics*, **20**, 1373–1380.

Andreini,C. *et al*. (2008) Metal ions in biological catalysis: from enzyme databases to general principles. *J. Biol. Inorg. Chem*., **13**, 1205–1218.

Andreini,C. *et al*. (2011) A simple protocol for the comparative analysis of the structure and occurrence of biochemical pathways across superkingdoms. *J. Chem. Inf. Model*., **51**, 730–738.

Andreini,C. *et al*. (2013) MetalPDB: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res*., **41**, D312–D319.

Andreini,C. *et al*. (2016) Exploiting bacterial operons to illuminate human iron–sulfur proteins. *J. Proteome Res*., **15**, 1308–1322.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Estellon,J. *et al*. (2014) An integrative computational model for large-scale identification of metalloproteins in microbial genomes: a focus on iron–sulfur cluster proteins. *Metallomics*., **6**, 1913–1930.

Gladyshev,V.N. and Zhang,Y. (2013) Comparative genomics analysis of the metallomes. *Met. Ions. Life Sci*., **12**, 529–580.

Lin,H.H. *et al*. (2006) Prediction of the functional class of metal-binding proteins from sequence derived physicochemical properties by support vector machine approach. *BMC Bioinformatics*, **7**, S13.

Passerini,A. *et al*. (2011) MetalDetector v2.0: predicting the geometry of metal binding sites from protein sequence. *Nucleic Acids Res*., **41**, W288–W292.

Paul,V.D. and Lill,R. (2015) Biogenesis of cytosolic and nuclear iron–sulfur proteins and their role in genome stability. *Biochim. Biophys. Acta*, **1853**, 1528–1539.